Studies in Systems, Decision and Control 164

Habib M. Ammari *Editor*

Mission-Oriented Sensor Networks and Systems: Art and Science

Volume 2: Advances

Springer

Studies in Systems, Decision and Control

Volume 164

Series Editor

Janusz Kacprzyk, Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

The series "Studies in Systems, Decision and Control" (SSDC) covers both new developments and advances, as well as the state of the art, in the various areas of broadly perceived systems, decision making and control–quickly, up to date and with a high quality. The intent is to cover the theory, applications, and perspectives on the state of the art and future developments relevant to systems, decision making, control, complex processes and related areas, as embedded in the fields of engineering, computer science, physics, economics, social and life sciences, as well as the paradigms and methodologies behind them. The series contains monographs, textbooks, lecture notes and edited volumes in systems, decision making and control spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

** Indexing: The books of this series are submitted to ISI, SCOPUS, DBLP, Ulrichs, MathSciNet, Current Mathematical Publications, Mathematical Reviews, Zentralblatt Math: MetaPress and Springerlink.

More information about this series at http://www.springer.com/series/13304

Habib M. Ammari
Editor

Mission-Oriented Sensor Networks and Systems: Art and Science

Volume 2: Advances

Springer

Editor
Habib M. Ammari
Wireless Sensor and Mobile Ad-hoc Network
Applied Cryptography Engineering
(WiSeMAN-ACE) Research Lab
Department of Electrical Engineering
and Computer Science
Frank H. Dotterweich College of Engineering
Texas A&M University-Kingsville
Kingsville, TX, USA

ISSN 2198-4182 ISSN 2198-4190 (electronic)
Studies in Systems, Decision and Control
ISBN 978-3-319-92383-3 ISBN 978-3-319-92384-0 (eBook)
https://doi.org/10.1007/978-3-319-92384-0

Library of Congress Control Number: 2018941995

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Allah, the Most Beneficent, the Most Merciful; and His Prophet, Mohamed, Peace Be Upon Him

To my first teachers: My mother, Mbarka, and my father, Mokhtar.

To my very best friends: My lovely wife Fadhila, and beautiful children, Leena, Muath, Mohamed-Eyed, Lama, and Maitham.

To my dearest brother, Lazhar, and sisters, Naima, Saloua, Monia, Faouzia, Alia, and Najeh.

To all my wonderful nieces, Rahma, Safa, Marwa, Amani, Omayma, and Sabrine; and my nephews, Mohamed, Bilel, Ahmed, and Youssef.

To the profound souls of my grandparents, Abdelkarim and Fatma, and my uncle, Mahfoudh.

To my Dean, Dr. Mohammad S. Alam, Fellow-IEEE, IET, OSA, SPIE, IoP, IS&T, and IAPR, Professor of Electrical Engineering, in the Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, at Texas A&M University-Kingsville, for his outstanding support to me since I joined Texas A&M University-Kingsville in August 2019.

To Dr. Afzel Noore, Associate Dean for Undergraduate Affairs, and Professor of Computer Science in the Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, at Texas A&M University-Kingsville, for his excellent support to me since I joined Texas A&M University-Kingsville in August 2019.

To the profound soul of my Professor, Dr. Hedi Ben Saad, Professor of Mathematics in the Department of Mathematics at the Faculty of Sciences of Tunis, Tunisia (May 25, 1950–June 30, 2016). He taught me mathematics in my second year of Physics and Chemistry major at the Faculty of Sciences of Tunis during the academic year 1987–1988. He was extremely knowledgeable in mathematics, very humble, and so kind. I have never seen anyone so far in his goodness, humility, and acute intelligence in mathematics.

To the profound soul of my Provost, Dr. Stephen Freedman (April 7, 1950–July 2, 2018), Professor of Biology and Senior Vice President for Academic Affairs and Chief Academic Officer at Fordham University, for his outstanding support to me during my stay at Fordham University.

To all of my friends and colleagues in the Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, at Texas A&M University-Kingsville, for their wonderful friendship and outstanding support to me since I joined Texas A&M University-Kingsville in August 2019.

Foreword

It has been more than a decade and a half since the first set of research papers on wireless sensor networks came out in the early 2000s. Academic research through the 2000s and 2010s has addressed many problems in the area rather comprehensively—from communication and routing protocols to time synchronization and localization, many problems that once appeared wide open have a number of practical and efficient solutions that can be used by practitioners.

Another transformation that has occurred in this time is that researchers have figured out how to place the problems of sensor networks within the broader context of large-scale cyber-physical systems and the Internet of Things. And entirely new areas of research have started to emerge, such as wireless charging, robotic wireless sensor networks, vehicular networks, and drone swarms.

Mission-Oriented Sensor Networks and Systems: Art and Science is a first-rate two-volume book with a collection of chapters contributed by experts, which provide not only a broad overview of many of the mature topics in this field but also a thorough introduction to some of the newer areas of inquiry. It is truly an outstanding contribution to the literature that owes much to the diligent efforts of Professor Habib M. Ammari from Texas A&M University-Kingsville, who has single-handedly edited both volumes.

As indicated in its title, a running theme through the book is a consideration of mission-critical applications where the dependability and security issues of sensor networks must be given significant attention. After an introduction, the book begins with an exploration of these mission-critical issues, from communication and routing perspectives, in Part I (Chapters "Autonomous Cooperative Routing for Mission-Critical Applications", and "Using Models for Communication in Cyber-Physical Systems").

Part II of the book explores ideas at the transition to the Internet of Things, with Chapters "Urban Microclimate Monitoring Using IoT-Based Architecture", "Models for Plug-and-Play IoT Architectures", and "Digital Forensics for IoT and WSNs" addressing topics from microclimate monitoring to localization, while these topics have been at the heart of sensor networking research since the beginning.

The shift toward IoT has highlighted the design of more ambitious applications often spanning large areas and the use of wide-area networks.

Another aspect of interest in this domain has been the development of crowd-sensing applications in the context of smart cities, treated in Part III. Developing appropriate incentives for users contributing data remains an area with significant open questions.

Part IV looks at wearable computing, which has been a trend on the rise for IoT systems intended to be able to collect relevant meaningful information about and from individuals that is relevant for many applications, while Part V of the book looks at novel approaches to wireless charging and how they can be integrated with real applications. Part VI of the book turns to providing an overview of robotic wireless sensor networks, which bring to the fore new challenges and opportunities associated with the added dimension of (often, decentrally) controlled mobility. Finally, last but not the least, the seventh and final part focuses on reliability, security, and interference mitigation.

These chapters collectively point to some of the most cutting-edge topics in sensor networks today. I believe the prospect is bright for further enhancements and developments, particularly as IoT and dependable cyber-physical systems become an increasingly bigger part of smart city operation.

As one example of ongoing academic work relevant to sensor networks, at the University of Southern California, together with my students and collaborators, I have been exploring the design of IoT data marketplaces for smart cities. In such marketplaces, buyers and sellers of data from IoT devices can interact with each other systematically over an economic layer. We have also been exploring how blockchain technologies can provide an additional layer of trust, robustness, and ultimately decentralization, compared to traditional approaches. And finally, we have been exploring how distributed computing and networked robotics systems can be developed from the ground-up.

Something we are likely to see going ahead in this field is a greater focus on systems rather than just algorithmic components or building blocks, and as these systems develop and need to scale in a heterogeneous and distributed manner, there will be new challenges such as the problem of interoperability and others. These will provide opportunities for corresponding solution techniques to be developed by researchers. Sensor networking as a whole is going to remain a relevant subject for years to come. This book is making a valuable and timely contribution in helping researchers catch up quickly with the latest advances in various aspects of the field.

Los Angeles, USA
October 25, 2018

Bhaskar Krishnamachari
Professor and Director
Center for Cyber-Physical Systems
and the Internet of Things
Viterbi School of Engineering
University of Southern California

Contents

About the Editor

Habib M. Ammari is an Associate Professor and the Founding Director of Wireless Sensor and Mobile Ad-hoc Network Applied Cryptography Engineering (WiSeMAN-ACE) Research Lab, in the Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, at Texas A&M University-Kingsville (TAMUK) since August 2019. He received his tenure in May 2014 in the Department of Computer and Information Science, College of Engineering and Computer Science, at the University of Michigan-Dearborn, where he served on the Distinguished Research Award Committee in 2015. Also, he received tenure at the Higher School of Communications in Tunis, Tunisia (Sup'Com Tunis) in 1998. Recently, he received the 2018 Albert Nelson Marquis Lifetime Achievement Award. He was selected as instructor at Stanford University in the Stanford Summer College Academy 2016 program, where he taught "Discrete Mathematical Structures: Foundational Concepts in Computer Science, Engineering, and Mathematics". He obtained his second Ph.D. degree in Computer Science and Engineering from the University of Texas at Arlington, in May 2008, and his first Ph.D. in Computer Science from the Faculty of Sciences of Tunis, in December 1996. His main research interests lay in the area of wireless sensor and mobile ad hoc networks, including connected *k*-coverage, geographic forwarding, physical and information security, applied cryptography, and computational geometry in wireless sensor networks. He has a strong publication record in top-quality journals, such as ACM TOSN, ACM TAAS,

IEEE TPDS, IEEE TC, Elsevier Ad Hoc Networks, Elsevier COMNET, Elsevier PMC, Elsevier JPDC, Elsevier COMCOM, and high-quality conferences, such as IEEE SECON, IEEE ICDCS, IEEE MASS, and EWSN. He published his first Springer book, "Challenges and Opportunities of Connected *k*-Covered Wireless Sensor Networks: From Sensor Deployment to Data Gathering" in August 2009. Also, he is the author and editor of two Springer books, "The Art of Wireless Sensor Networks: Fundamentals" and "The Art of Wireless Sensor Networks: Advanced Topics and Applications", which have been published in January 2014. In addition, he published these two current Springer books, "Mission-oriented sensor networks and systems: Art and science–Foundations" and "Mission-oriented sensor networks and systems: Art and science–Advances" in January 2019. He is the recipient of the US National Science Foundation (NSF) CAREER Award in January 2011, a 3-year US NSF Research Grant Award in June 2009, the National Security Agency (NSA) Award in 2017, and the Faculty Research and Development Grant Award from Hofstra College of Liberal Arts and Sciences in May 2009. In March 2014, he was recognized with the Distinguished Research Award at the University of Michigan-Dearborn. Furthermore, in May 2010, he was recognized with the Lawrence A. Stessin Prize for Outstanding Scholarly Publication (*i.e.,* Distinguished Research Award) at Hofstra University. He is the recipient of the Nortel Outstanding CSE Doctoral Dissertation Award in February 2009, and the John Steven Schuchman Award for 2006–2007 Outstanding Research by a Ph. D. student in February 2008. He received the Best Paper Award at EWSN in 2008, and the Best Paper Award at the IEEE PerCom 2008 Google Ph.D. Forum. He received several other prestigious awards, including the Best Graduate Student Paper Award (Nokia Budding Wireless Innovators Awards First Prize) in May 2004, the Best Graduate Student Presentation Award (Ericsson Award First Prize) in February 2004, and Laureate in Physics and Chemistry for academic years 1987 and 1988. Also, he was selected as the ACM Student Research Competition Finalist at the ACM MobiCom in September 2005. Also, he was selected for inclusion in the Marquis Who's Who in the World in 2019 and 2018,

AcademicKeys Who's Who in Sciences Higher Education in 2017, Who's Who in America in 2017, AcademicKeys Who's Who in Engineering Higher Education in 2012, the AcademicKeys Who's Who in Sciences Higher Education in 2011, Feature Alumnus in the University of Texas at Arlington CSE Department's Newsletter in Spring 2011, Who's Who in America in 2010, and the 2008-2009 Honors Edition of Madison Who's Who Among Executives and Professionals. He received several service awards, including the Certificate of Appreciation Award at MiSeNet 2014, the Certificate of Appreciation Award at ACM MiSeNet 2013, the Certificate of Appreciation Award at the IEEE DCoSS 2013, the Certificate of Appreciation Award at the ACM MobiCom 2011, the Outstanding Leadership Award at the IEEE ICCCN 2011, and the Best Symposium Award at the IEEE IWCMC 2011. He serves as the Founding Coordinator of the CIS Distinguished Lecture Series, and as Coordinator of the CIS Faculty Research Talk Series since 2017. In addition, he was the Founding Coordinator of both the Distinguished Lecture Series and the Research Colloquium Series, in the College of Engineering and Computer Science at the University of Michigan-Dearborn from 2011 to 2015. He was successful to invite ACM Turing Award Winners to his distinguished lecture series, such as Dr. Manuel Blum from Carnegie Mellon University (CMU), and Dr. Shafi Goldwasser from MIT, who gave talks at the University of Michigan-Dearborn on January 25, 2013, and October 25, 2013, respectively, and Dr. Martin E. Hellman from Stanford University, who gave a talk at Fordham University on October 22, 2018. He was invited to give several invited talks at reputed universities. Indeed, he was invited to give a talk at the Third Arab-American Frontiers of Sensor Science Symposium, which was organized by the US National Academy of Sciences on December 5–7, 2015. Also, he served as external examiner of several Ph.D. Dissertations. He is the Founder of the Annual International Workshop on Mission-Oriented Wireless Sensor Networking (MiSeNet), which has been co-located with ACM MobiCom, IEEE INFOCOM, and IEEE MASS conferences since 2012. He served as Associate Editor of several prestigious journals, such as ACM TOSN, IEEE TC, IEEE Access, and Elsevier PMC. He serves

on the Steering Committee of MiSeNet, the Annual International Conference on Distributed Computing in Sensor Systems (DCOSS), and the International Workshop on Wireless Mesh and Ad-hoc Networking (WiMAN). Moreover, he served as General Chair, Program Chair, Track Chair, Session Chair, Publicity Chair, Web Chair, and Technical Program Committee member of numerous ACM and IEEE conferences, symposia, and workshops. He is an IEEE Senior Member.

Introduction

Habib M. Ammari

Imagination is more important than knowledge. For knowledge is limited, whereas imagination embraces the entire world, stimulating progress, giving birth to evolution.

Albert Einstein, (1879–1955).

1 Mission-Oriented Sensor Networks and Systems: Art and Science

The fast advances in both inexpensive sensor technology and wireless communications over the last two decades have made the design and development of large-scale wireless sensor networks cost-effective and appealing to a wide range of mission-critical situations. These include area monitoring (e.g., deploying sensors for enemy intrusion detection, as well as geo-fencing of gas, oil pipelines, or work area), healthcare monitoring (e.g., using implanted, wearable, or environment-embedded sensors for medical applications), environmental/earth sensing (e.g., using sensors for monitoring air pollution and water quality, as well as detecting forest fire and landslide), industrial monitoring (e.g., deploying sensors for monitoring machine health, data center, data logging, water and wastewater, and structural health), to name a few.

Wireless sensor networking has attracted the attention of numerous practitioners and researchers from both industry and academia. These types of networks consist of

H. M. Ammari (✉)
Wireless Sensor and Mobile Ad-hoc Network Applied Cryptography Engineering (WiSeMAN-ACE) Research Lab, Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, Texas A&M University-Kingsville, Kingsville, TX, USA
e-mail: hammari@fordham.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_1

a collection of tiny, resource-limited, low-reliable sensing devices that are randomly or deterministically deployed in a field of interest to monitor a physical phenomenon and report their results to a central gathering point, known as *sink* (or *base station*). These tiny sensing devices suffer from their scarce capabilities, such as bandwidth, storage, CPU, battery power (or energy), sensing, and communication. In particular, the constrained power supplies of the sensors shorten their lifetime and make them unreliable. More precisely, mission-oriented sensor networks and systems are viewed as time-varying systems composed of autonomous (mobile) sensing devices (e.g., using mobile robots) that collaborate and coordinate distributedly to successfully accomplish complex real-time missions under uncertainty. The major challenge in the design of mission-oriented sensor networks and systems is due to their dynamic topology and architecture, which is caused mainly by sensing devices mobility. The latter may have significant impact on the performance of mission-oriented sensor networks and systems in terms of their sensing coverage and network connectivity. In such continuously dynamic environments, sensing devices should self-organize and move purposefully to accomplish any mission in their deployment field while extending the operational network lifetime. In particular, the design of mission-oriented sensor networks and systems should account for trade-offs between several attributes, such as energy consumption (due to mobility, sensing, and communication), reliability, fault tolerance, and delay.

Mission-oriented sensor networks and systems have been able to attract the attention of numerous people from scientific communities in both academia and industry. Indeed, a large number of related innovative research papers to solve challenging problems have been published in high-quality journals, conferences, and workshops. Given the importance of this area of research, I found it is essential that an up-to-date book on the abovementioned topics be provided to our sensor networks and systems research community. This book series, titled "*Mission-Oriented Sensor Networks and Systems: Art and Science*", includes two volumes, namely, Volume 1 and Volume 2, whose titles are as follows, respectively,

- *Mission-Oriented Sensor Networks and Systems : Art and Science—Foundations*
- *Mission-Oriented Sensor Networks and Systems : Art and Science—Advances*

These two books have been assembled with a goal to address challenging and/or open research problems in traditional as well as new emerging areas of research in mission-oriented sensor networks and systems, including sensor networking, cyber-physical systems, and Internet of things, to name a few. It is worth mentioning that all the book chapters in both volumes have been written as comprehensive review of the state-of-the-art and state-of-the-practice of their associated topics. Precisely, each book chapter is either a survey of existing work in the literature, or a survey with emphasis on the related research done by their corresponding authors. In either case, every book chapter presents a thorough review of the underlying theoretical foundations, along with in-depth overview of the proposed approaches.

This book relates to the second volume, i.e., *Mission-Oriented Sensor Networks and Systems : Art and Science—Advances*. It focuses on advances in mission-oriented sensor networks and systems, including nonconventional sensor networks and sys-

tems, which are not covered in the first volume. Thus, the second volume of this series deals with advanced topics in mission-oriented sensor networks and systems, such as cyber-physical systems, Internet of things (IoT), crowdsensing, wearable computing, robotics, and wireless charging systems. For instance, cyber-physical systems (CPSs) have emerged as a promising technology to provide a bridge between the physical and cyber worlds, where sensors, actuators, and embedded devices are networked to sense, monitor, and control the physical world. Several applications of CPSs, such as health care, transportation, and rescue applications, have been developed. Note that all of these CPSs applications leverage the data gathered by sensors in order to bridge between the physical and cyber worlds. Following Albert Einstein's above-quoted-wise approach of being imaginative, all the book chapters in this second volume emphasize real systems that can be designed mainly based on sensors. These book chapters include up-to-date research work spanning various topics in mission-oriented sensor networks and systems, such as autonomous cooperative routing for mission-critical applications, communication of models and model updates as new paradigm in communication, IoT-based architecture, models for IoT architectures, digital forensics, dependable wireless communication in IoT, localization in IoT, mobile crowdsensing systems, smart cities, privacy and security in wearable computing, wireless transfer of energy, robotics, middleware for robotics, interference mitigation, radiation control, and encryption. Most of these major topics can be covered in an advanced course on mission-oriented sensor networks and systems. I believe that this book will be an excellent reference for graduate as well as senior undergraduate students who are majoring in computer science, computer engineering, electrical engineering, data science, information science, or any related discipline. Furthermore, this book will a great source of information for computer scientists, researchers, and practitioners in academia and industry. I really hope that all readers will find this book very useful, nicely written, clear, exciting, and fascinating. My ultimate goal is that all users of this book will enjoy reading it and using it for any of their favorite research topics, as much as I enjoyed editing it.

2 Book Organization

This book consists of seven parts, each of which has two to four chapters. Next, we provide a short description of each part through a brief summary of each of its chapters.

In Part 1, titled "*Mission-Critical Applications and Cyber-Physical Systems*", Chapter "Autonomous Cooperative Routing for Mission-Critical Applications" discusses autonomous cooperative routing for mission-critical applications. It presents the underlying networking challenges and practical remedies for this type of routing. Also, it shows that autonomous cooperative routing outperforms other routing schemes. Chapter "Using Models for Communication in Cyber-Physical Systems" describes a new paradigm in communication, which utilizes communication of models and model updates instead of raw sensed data. Then, it demonstrates the effectiveness of the model-based communication using the example of a vehicular cyber-physical system.

In Part 2, titled "*Internet of Things*", Chapter "Urban Microclimate Monitoring Using IoT-Based Architecture" presents various aspects related to Internet of Things-based sensor node development for urban microclimate monitoring. It emphasizes software development, relevant methodologies, hardware modules, and platforms. Chapter "Models for Plug-and-Play IoT Architectures" reviews plug-and-play architecture along with its corresponding components. Then, it presents a survey of the most important models that feature the capabilities of this type of architecture. Chapter "Digital Forensics for IoT and WSNs" describes digital forensics challenges in the Internet of Things and wireless sensor networks environments. Also, it analyzes available solutions to overcome some of those challenges from different perspectives. Chapter "Dependable Wireless Communication and Localization in the Internet of Things" surveys methods to increase the dependability of the Internet of Things. It provides a comprehensive treatment of dependability issues across multiple layers, from signal processing, over microwave engineering, to networking. Also, it proposes to use a switchable ultra-wideband antenna system to enhance communication and localization in the Internet of Things. Moreover, it shows its potential for multipath-resolved positioning.

In Part 3, titled "*Crowdsensing and Smart Cities*", Chapter "User Incentivization in Mobile Crowdsensing Systems" identifies basic design issues of mobile crowdsensing systems and investigates some characteristic challenges. Chapter "Vehicular Ad Hoc/Sensor Networks in Smart Cities" introduces vehicular networks and their challenges. Then, it discusses some existing routing protocols for vehicular networks. Also, it describes some vehicular sensor applications in smart cities.

In Part 4, titled "*Wearable Computing*", Chapter "An Overview of Wearable Computing" provides a high-level user-oriented overview of wearable computing. Also, it gives a historical view of wearable computing devices, beginning with an abacus ring from the seventeenth century and progressing to modern wearable computing devices. Then, it discusses the lessons learned from this history and their implications for future wearable devices. Chapter "Wearables Security and Privacy" discusses security and privacy problems with wearable devices. It describes the components in wearables, such as sensors, processors, software, and communications, and highlights the security issues related to wireless protocols, vulnerabilities, and privacy. Chapter "Wearable Computing and Human-Centricity" focuses on the principles of human-centric design, which have been used in the context of wearable computers. Those traditional concepts of human-centric computers need to be defined in an adaptable framework. Indeed, wearable devices introduce an additional set of requirements to those traditional concepts. This chapter shows that human centricity is one of the major challenges to the ubiquity and future success of wearable devices.

In Part 5, titled "*Wireless Charging and Energy Transfer*", Chapter "Wireless Transfer of Energy Alongside Information in Wireless Sensor Networks" explores techniques to simultaneously deliver energy alongside information during wireless communications. It presents mechanisms to consolidate energy and information transfer in wireless sensor networks. Also, provides an experimental evaluation of the proposed iPoint communication system, which includes novel communication protocols and optimization techniques to ensure efficient delivery of energy and information. Chapter "Efficient Protocols for Peer-to-Peer Wireless Power Transfer and Energy-Aware Network Formation" investigates interactive, "peer-to-peer" wireless

energy exchange in populations of resource-limited mobile agents, without the use of any special chargers. Then, it discusses protocols that address energy balancing between agents, and distributively forming a network structure with an appropriate energy distribution among the agents. Chapter "Next-Generation Software-Defined Wireless Charging System" introduces an architecture for next-generation wireless charging systems, called DeepCharge, which realizes a software-defined wireless charging system through separation of controller, energy, and hardware planes. Then, it demonstrates indoor and outdoor prototypes of DeepCharge with extensive experimental measurement.

In Part 6, titled "*Robotics and Middleware*", Chapter "Robotic Wireless Sensor Networks" presents a literature survey of an emerging, cutting-edge, and multidisciplinary field of research in robotic wireless sensor networks. It identifies the core problems, such as connectivity, localization, routing, and robust flow of information. Then, it classifies the existing research on robotic wireless sensor networks. Also, it analyzes what is missing in the literature and identifies topics for future research. Chapter "Robot and Drone Localization in GPS-Denied Areas" discusses many facets of robot and drone coordination in GPS-denied areas. Also, it addresses issues associated with localization and coordination among multiple agents to accomplish a common goal. Chapter "Middleware for Multi-robot Systems" surveys state-of-the-art in both distributed multi-robot system and middleware. Then, it provides a taxonomy to classify the MRS middleware and analyze existing middleware functionalities and features.

In Part 7, titled "*Interference Mitigation, Radiation Control, and Encryption*", Chapter "Interference Mitigation Techniques in Wireless Body Area Networks" analyzes the issues related to the coexistence amongst wireless body area networks (WBANs) and between WBANs and other wireless networks. Also, it provides a comparative review of the radio co-channel interference mitigation and avoidance techniques that exists in the literature. Chapter "Radiation Control Algorithms in Wireless Networks" focuses on two problems. The first problem, called minimum radiation path, consists to find the lowest radiation trajectory of a person moving from a source to a destination point within the area of a network of wireless devices. The second problem is to efficiently charge a set of rechargeable nodes using a set of wireless energy chargers, under safety constraints on the electromagnetic radiation incurred. Then, it presents and analyzes efficient algorithms and heuristics for approximating optimal solutions, namely, minimum radiation trajectories and charging schemes, for both problems, respectively. Chapter "Subspace-Based Encryption" shows the weaknesses from a cryptographic point of view of the concept of blind source separation, which has been used for speech encryption. It proposes to use vectorial subspace concepts, leading to subspace-based encryption systems, which are applied to speech and images. Also, it shows through experiments that the use of subspace-based encryption systems yields performance enhancement.

3 Acknowledgements

This complete two-volume series book, titled "*Mission-Oriented Sensor Networks and Systems: Art and Science*", is a tribute to the outstanding work of the foremost

leading authorities and scholars in their fields of research in the area of mission-oriented sensor networks and systems. Honestly, it is unfair that my name only appears on the book cover. And, it is really a great pleasure and an honor for me to cordially recognize all of those who contributed a lot to this book and generously supported me throughout this project in order to make this two-volume series a reality. Therefore, it is really a great privilege for me to work with all of these talented scholars. Without them, it would not be possible at all to finish this book and make it available to all the researchers and practitioners, who are interested in the foundations of mission-oriented sensor networks and systems.

First and foremost, I am sincerely and permanently grateful to Allah—the Most Gracious, the Most Merciful—for everything He has been giving me. In particular, I would very much love to thank Him for providing me the golden opportunity to work with such group of outstanding scientists and researchers to put together this book, and for helping me publish it within 3 years. I am extremely happy and so excited to dedicate this modest book to Him, and very much hope that He would kindly accept it and put His Blessings in it. His Saying "**And of knowledge, you (mankind) have been given only a little**" has an endless, pleasant echo in my heart and always reminds me that our knowledge is much less than a drop in the ocean.

It is worth mentioning that all the contributing authors were invited to contribute to this book, and that no Call for Book Chapters had ever been sent out through any mailing list. All of those authors whom I invited were chosen very selectively to cover most of the foundational topics in mission-oriented sensor networks and systems. They have been contributing to the growth and development of the field of mission-oriented sensor networks and systems. This book would never have been written without their great contributions, support, and cooperation. Thus, my cordial recognition is due to all of my friends and colleagues (faculty and—the ones whom I invited to contribute with their book chapters to this book—whose names are listed in the alphabetical order: Drs. Kemal Akkaya, Mohamad Jaafar Ali, Mohamed-Slim Alouini, Constantinos Marios Angelopoulos, Ahmed Bader, Mustafa S. Bakr, Md Zakirul Alam Bhuiyan, Adel Belouchrani, Jorge Blasco, Carlo Alberto Boano, Wolfgang Bosch, Jiannong Cao, Thomas M. Chen, Kaushik R. Chowdhury, Yaser Fallah, Andrea Gasparri, Fabrizio Gentili, Pradipta Ghosh, Philipp Greiner, Jasmin Grosinger, Bernhard Großwindhager, Hooman Javaheri, Mukesh Jha, Shan Jiang, Jiong Jin, Umit Karabiyik, Bhaskar Krishnamachari, Adelina Madhja, Prashanth Reddy Marpu, Troy McDaniel, Ahmed Mehaoua, Atef Mermoul, Hassine Moungla, Ufuk Muncuk, M. Yousof Naderi, Sotiris Nikoletseas, Guevara Noubir, Sethuraman Panchanathan, Harsh Kupwade Patil, Christian Poellabauer, Theofanis P. Raptis, Christoforos Raptopoulos, Michael Rath, Jose Rolim, Kay Römer, Yuvraj Sahni, Josh Siva, Chao Song, Arash Tadayon, Ramin Tadayon, Dimitrios Tsolovos, Alexandros Tsoupos, Gary M. Weiss, Klaus Witrisal, Daniel Wolff, Jie Wu, and Mohamed Younis. I am really honored to have worked with such an amazing crew of scholars and scientists. I learned a lot from them throughout this project, and it was an incredible experience for me in finishing this book.

Each book chapter has undergone two rounds of reviews. Moreover, in each round, every book chapter received 2–5 reviews by experts in the scope of the chapter. Our ultimate goal is to provide the readers with a high-quality reference on the founda-

tions of mission-oriented sensor networks and systems. Precisely, all book chapters were carefully reviewed in both rounds by all the contributing authors. I would like to express my sincere gratitude to all the contributing authors for their constructive feedback to improve the organization and content of all book chapters. My special thanks go to Drs. Damian M. Lyons (external reviewer), Flavia Delicato, and Mohamed Younis for their generous offer to review several book chapters of both books of this two-volume series. Also, my original plan was to publish only one book, titled "*Mission-Oriented Sensor Networks and Systems: Art and Science*". But, I ended up with 42 book chapters, which I split into two volumes along with their book chapters and titles. Moreover, I am very grateful to Dr. Bhaskar Krishnamachari, Professor and Director, Center for Cyber-Physical Systems and the Internet of Things Viterbi School of Engineering, University of Southern California, Los Angeles, California, for his great foreword, kindness, and outstanding support to me.

I started this project on Monday, September 5, 2016 at 12:42 AM when I contacted the Publishing Editor, Dr. Thomas Ditzinger, who approved my proposal for an edited book. All book chapters for both volumes were uploaded on the website of Springer and made accessible to the Project Coordinator, Mr. Gowrishankar Ayyasamy, on August 21, 2019. Hence, this project lasted about 3 years. During all this period of time, I exchanged a few thousands of emails with all contributing authors with regard to their book chapters. I would like to thank all the contributing authors for their invaluable time, flexibility, and wonderful patience in responding to all of my emails in a timely manner. Please forgive me for your time, and I hope that the readers will appreciate all of your great efforts and love all the materials in this book. We all have devoted a considerable amount of time to finish this book, and I hope that all of our efforts will be paid off in the future.

I would like to acknowledge all of my family members who have provided me with excellent source of support and constant encouragement over the course of this project. First of all, I am extremely grateful to both of my first teachers, my mother, Mbarka, and my father, Mokhtar, for their sincere prayers, love, support, and encouragement, and for always teaching me and reminding me of the value of knowledge and the importance of family. I owe them a lot, and cannot find my words to thank them enough for everything they have done to make me who I am now. Also, I am most grateful to my best friend and beloved wife, Fadhila, for her genuine friendship and for being extremely supportive and unboundedly patient while I was working on this book. In addition, I would like to express my hearty gratitude to my lovely and beautiful children, Leena, Muath, Mohamed-Eyed, Lama, and Maitham, for their endless love, support, and encouragement. They have been one of my greatest joys, very patient, and understanding. I hope they will forgive me for spending several hours away from them, while I was setting in front of my PC in my office or my laptop at home busy with this book. Several times, they all told me: "Daddy, as usual, your books and emails are always dragging you away from us!" My lovely wife and children have been a wonderful inspiration to me, and very patient throughout the life of this project. Without their warm love and care, this project would never even have been started. Furthermore, my special thanks and gratitude go to all of my sisters, brother, nieces, and nephews for their love, thoughtful prayers, concern, and valuable support all the time.

This project could not have been completed without the great support of the people around me who made this experience successful and more than enjoyable. I would like to thank all of my friends and colleagues at Texas A&M University-Kingsville, and, particularly, my fellows in the Department of Electrical Engineering and Computer Science, for the collegial and very friendly atmosphere they provided me with to finish this book. In particular, I am very grateful to my Dean, Dr. Mohammad Alam, Fellow-IEEE, IET, OSA, SPIE, IoP, IS&T, and IAPR, Professor of Electrical Engineering, in the Department of Electrical Engineering and Computer Science, Frank H. Dotterweich College of Engineering, at Texas A&M University-Kingsville, for his kindness, continuous encouragement, and outstanding support to WiSeMAN-ACE Research Lab since I joined the Department of Electrical Engineering and Computer Science at Texas A&M University-Kingsville in August 2019. Also, I am very thankful to Dr. Afzel Noore, Professor and Associate Dean for Undergraduate Affairs in the Frank H. Dotterweich College of Engineering at Texas A&M University-Kingsville, for his humbleness and outstanding support to me in several ways. Furthermore, I would like to express my profound gratitude to all of my colleagues, including faculty and staff members, in the Department of Electrical Engineering and Computer Science at Texas A&M University-Kingsville, especially Drs. Rajab Challoo, Reza Nekovei, Syed Iqbal Omar, Sung-won Park, Scott Smith (EECS Department Chair), Lifford Mclauchlan, Mais Nijim, Amit Verma, Muhittin Yilmaz, Nuri Yilmazer, Muhammad Aurangzeb, Gahangir Hossain, and Maleq Khan; and Mrs. Debra Beltran and Mr. G.R. Benavides, for their extended support and encouragement. Moreover, I would like to convey my warm thanks and appreciation to all the faculty and staff members in the Frank H. Dotterweich College of Engineering Dean's Office who have been so helpful and very kind to me, namely, Drs. Mahesh Hosur, Associate Dean for Graduate Affairs and Research, and Robert Diersing, Executive Director, High-Performance Computing Center; and Tamara Denise Guillen, Julissa Flores, Rosenda Garcia, and Rose Anna Gomez. Also, I would like to convey my special thanks and deep appreciation to Dr. Zakaria Abd-Elmageed, Professor in the College of Pharmacy at Texas A&M University, College Station, for his friendship, kindness, and outstanding support to me in several ways. This work is partially supported by the National Science Foundation (NSF) grants 0917089 and 1054935.

Last, but not the least, I would like to express my deep appreciation and gratitude to Dr. Thomas Ditzinger, Publishing Editor; Mr. Gowrishankar Ayyasamy and Ms. Janet Sterritt-Brunner, Project Coordinators for Books Production; Ms. Daniela Brandt, Project Coordinator for Publishing; Ms. Sabine Gutfleisch, Editorial Assistant; Mrs. Sudhany Karthick, Project Manager for Book Production; and Ms. Sylvia Schneider, Production Coordinator for Books Production. It was a great pleasure to work with all of them. I would like to acknowledge the publisher, Springer, for the professionalism, patience, and the high quality of their typesetting team as well as their timely publication of this book.

August 21, 2019

Part I
Mission-Critical Applications and Cyber-Physical Systems

Autonomous Cooperative Routing for Mission-Critical Applications

Ahmed Bader and Mohamed-Slim Alouini

Abstract We are entering an era where three previously decoupled domains of technology are rapidly converging together: robotics and wireless communications. We have seen giant leaps and improvements in computational efficiency of vision processing and sensing circuitry coupled with continuously miniaturized form factors. As a result, a new wave of mission-critical systems has been unleashed in fields like emergency response, public safety, law enforcement, search and rescue, as well as industrial asset mapping. There is growing evidence showing that the efficacy of team-based mission-critical systems is substantially improved when situational awareness data, such as real-time video, is disseminated within the network. Field commanders or operation managers can make great use of live vision feeds to make educated decisions in the face of unfolding circumstances or events. In the likely absence of adequate cellular service, this translates into the need for a mobile ad hoc networking technology (MANET) that supports high throughput but more importantly low end-to-end latency. However, classical MANET technologies fall short in terms of scalability, bandwidth, and latency; all three metrics being quite essential for mission-critical applications. The real bottleneck has always been in how fast packets can be routed through the network. To that end, autonomous cooperative routing (ACR) has gained traction as the most viable MANET routing proposition. Compared to classical MANET routing schemes, ACR is poised to offer up to 2X better throughput, more than 4X reduction in end-to-end latency, while observing a given target of transport rate normalized to energy consumption. Nonetheless, ACR is also associated with a few practical implementation challenges. If these go unaddressed, it will deem ACR practically infeasible. In this chapter, efficient and low-complexity remedies to those issues are presented, analyzed, and validated. The

A. Bader · M.-S. Alouini (✉)
Computer, Electrical, and Mathematical Science and Engineering (CEMSE) Division,
King Abdullah University of Science and Technology (KAUST),
Thuwal, Makkah Province, Saudi Arabia
e-mail: slim.alouini@kaust.edu.sa

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_2

validation is based on field experiments carried out using software-defined radio (SDR) platforms. This chapter sheds light on the underlying networking challenges and practical remedies for ACR to fulfill its promise.

Keywords Autonomous cooperative routing
Mobile ad hoc networks (MANET) · Mission-critical applications
Situational awareness · Real-time video streaming · Path-oriented routing
Geographical routing · End-to-end latency · Normalized transport rate
Cooperative transmission · Carrier frequency offset (CFO)
Software-defined radio (SDR)

1 Introduction

1.1 Team-Based Mission-Critical Applications

Broadly speaking, mission-critical applications are defined as those applications demanding data delivery bounds in the time and reliability domains [1]. When a critical mission is executed by a cluster or swarm of a human and/or robotic agents it is typically referred to as a team-based mission-critical operation [2]. In the abstract sense of things, a mission-critical agent is only able to execute its mission when equipped with the right sensory and possibly actuation gear. Thus, mission-oriented wireless sensor networks (the core subject of this book) and team-based mission-critical operations clearly intersect.

Recently, there has been an unprecedented growth in the use of computer vision in mission-critical applications [3]. The dissemination of live vision-based data feeds offers great visibility into the underlying process being monitored, mapped, or controlled [4, 5]. Live video streaming is believed to offer significant improvement in the decision-making abilities in the face of unexpected events [6]. As a matter of fact, the use of real-time video and vision-based data streaming for enhancing the contextual awareness levels has been lately earning substantial interest in other relevant domains such as telemedicine [7], paramedics [8], emergency and first response [9], law enforcement, and tactical (military) operations [10].

The availability of live video feeds is crucial in boosting situational and contextual awareness. It is argued that human decision-making failures during time- and mission-critical scenarios can be caused by shortage of understanding of the underlying situation and inability to understand the context [6]. Field commanders are consistently required to take decisions in response to events occurring in the field. There is appreciable evidence that acquiring access to live video feeds streamed from front-end personnel diminishes uncertainty and therefore upgrades the decision-making quality [9, 11]. It is not only raw-format video that proves to be useful, but other formats can even have more utility such as thermal vision data during hydrocarbon leak detection for instance.

The virtue of real-time vision-based data sharing in boosting the operational efficiency of mission-critical operations is hopefully quite intuitive. Human agents can make better decisions when offered timely information and deeper visibility into the ongoing physical process being treated [12]. Research has also shown that the availability of real-time video communications for paramedics and emergency responders significantly enhances collaborative execution of a mission and reduces time to completion [11]. Therefore, we have seen more emphasis on real-time video streaming for mission-critical operations in literature. For example, the U.S. National Institute of Standards and Technology (NIST) has lately released a technical note [13] in which the significance of real-time streaming for public safety operations is clearly underscored.

Real-time streaming from the field can be well extended to additional use cases in other industrial verticals. Two general categories of mission-critical applications are addressed herewith:

(1) Defense and emergency response operations mainly encompassing tactical missions, law enforcement, firefighting, search and rescue, crowd management, and telemedicine.
(2) Industrial field operations mainly in hydrocarbon exploration and production (E&P), mining, and power generation. Within this context, it is often required to dispatch crews of technicians and engineers to the field to execute a certain time-critical maintenance routine, react to a process failure, or treat a chemical spill.

In both categories, front-end field personnel are equipped with sensory that feeds back critical information to the back-end decision-making central. Such information is analyzed manually and/or automatically before commands and actuation instructions are fed forward to the front-end.

An example of a public safety mission-critical operation is illustrated in Fig. 1. From a networking viewpoint, data flow in mission-critical applications is mainly dictated by the underlying decision-making mechanism. Teams deployed into the field typically follow a hierarchical command chain [6]. Attempting to process or even just view the data by front-end personnel may cause distraction. Hence, decision-making by far is largely concentrated at the back-end point. This implies that the network has to generally operated according to a "convergecast" rather than peer-to-peer mode.

Figure 1 also showcases a growing trend toward future mission-critical MANET. It is envisioned that unmanned autonomous vehicles (UAV) will be deployed as front-end agents [2]. Here, swarms of aerial or terrestrial UAVs are dispatched into the field to execute a mission under human supervision and control. In mission-critical applications, a paramount task is the joint planning and optimization of motion trajectories of the human and robotic agents [14]. The timeliness of disseminating path planning and control signaling messages is quite instrumental [2]. This places

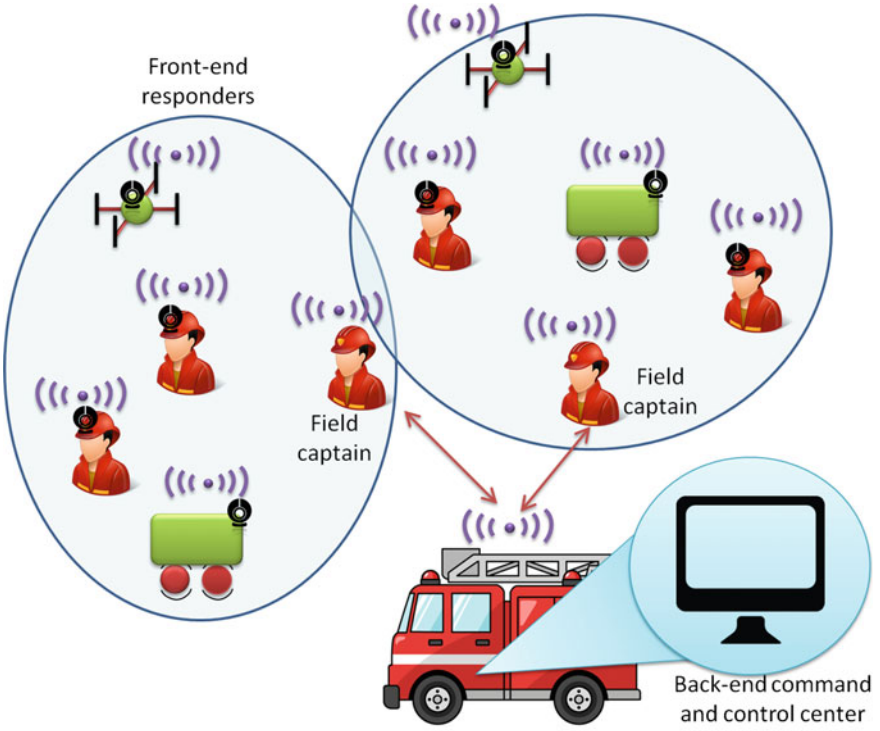


Fig. 1 An example of public safety MANET use case. Situational awareness data (video, thermal imaging, pressure, and temperature readings) flows toward the field commander and eventually to the back-end command and control center. Commands and actuation instructions are fed forward to front-end personnel as well as unmanned autonomous vehicles (UAV). Actuation instructions may include opening/closing a gate, controlling a valve, spraying of chemicals, etc.

yet an additional level of significance for designing low-latency MANET routing schemes.

The introduction of UAVs is also expected to go mainstream in mission-critical industrial operations. One example is the use of UAV swarms for thermal imaging and remote sensing [15]. UAVs can be deployed in industrial facilities as a routine maintenance measure, or as part of an emergency response operation. It is also noteworthy to mention at this point that UAV swarms have been also considered for 3D mapping, surveying, and other civil engineering tasks [16]. Again, the value proposition of deploying UAVs alongside human agents is manifested in the reduction of time to complete a mission, reduce injury rates, achieve better field coverage, and improving accessibility into hard-to-reach spots.

2 Mobile Ad Hoc Routing Revisited

By nature, real-time video streaming applications are typically delay-intolerant. Needless to say that vision-based data is also bandwidth-hungry. From a wireless networking perspective, it is indeed always desirable to capitalize to the maximum extent possible on the economies of scale offered by off-the-shelf standardized technologies. Hence, the natural technology candidates are LTE and Wi-Fi. Nonetheless, there are unfortunately some inherent deficiencies in LTE and Wi-Fi systems which render them less attractive, particularly for applications of mission-critical nature.

LTE as a cellular technology is ubiquitous but only to a limited extent. It is straightforward to argue that there will be situations and circumstances where adequate LTE service is not available [17]. Examples include remote onshore industrialized sites, offshore oil rigs, or deep mining pits. In fact, even in urbanized areas where coverage *does* exist, field personnel may have to be deployed in hard-to-reach areas where LTE does not penetrate indoors deeply enough.

Another interesting example where LTE is highly likely to fall short is massively crowded events [18]. In such contexts, the sheer scale of the load that LTE networks have to withstand has an adverse effect on the bandwidth and delay performance for mission-critical applications. One may argue that mission-critical applications are typically granted preemption on the radio access network (RAN) interface by mobile operators [19]. This is a valid argument so long as the mission-critical user equipment (UE) has already managed to gain access to the network. However, gaining access to the RAN in the first place may suffer from tremendous latencies and escalated rates of failure [20]. This specifically true under high user/traffic intensities; something which is quite expected in massively crowded events. The same rationale also applies during times of natural disasters when attempts to place calls on the network throttles the network.

Unlike LTE, Wi-Fi is more of a portable technology. This is actually meant in the sense that Wi-Fi hotspots[1] can be deployed by field personnel right in the area where action is taking place. However, due to regulatory and inherent design constraints, Wi-Fi only offers a limited reach when deployed as a single-hop network.

Attempts to extend Wi-Fi service coverage span can be accomplished by means of multihop networking. However, real-world deployments have repeatedly reported some hard limits on the number of hops that Wi-Fi-based solutions can sustain [21]. This is in part due to the excessive medium access control (MAC) layer overhead plaguing Wi-Fi. As a matter of fact, the IEEE 802.11ax standard (expected to be released in 2019) is already working on means to streamline the MAC and reducing the mean time to accessing the medium [22].

Having said that, the underlying MAC layer in Wi-Fi is not fully to blame. A major contributor to the non-scalability of Wi-Fi in multihop contexts is the routing overhead. This has been coined by some researchers as a cause of "capacity deficit" [23] and recognized as a major challenge by the Defense Advanced Research Projects

[1] LTE-Unlicensed hotspots is obviously a very comparable option to Wi-Fi. In other words, it will suffer more or less from the same scalability issues outlined herewith for the case of Wi-Fi.

Agency (DARPA) [24]. Such a deficit or shortcoming tends to have a more profound effect as the scale (number of users and/or traffic intensity) increases as well as with increased mobility.

As a result, there is an obvious need for infrastructure-independent ad hoc networking with strong support for mobility. Clearly, this can be articulated as a quest for a high-throughput low-latency mobile ad hoc network (MANET). Consequently, proprietary tailor-made MANET technologies are resurfacing again as viable propositions for mission-critical operations [25].

Undoubtedly, multihop MANET research literature has a mature legacy of work that is at least a couple of decades old. However, the need for significantly more bandwidth per user, ultralow end-to-end (e2e) latency, and tangibly better scalability calls for going back to the drawing board [26]. This is true since classical routing schemes are plagued by protocol overheads which have the tendency to substantially throttle the end-to-end performance of the MANET [27, 28].

To alleviate the routing overhead problem, autonomous cooperative routing (ACR) comes to rescue. In ACR, routing decisions are taking locally, i.e., wireless nodes do not revert to cross-coordination between each other before a packet is forwarded [29]. In fact, ACR does not revert to the classical concept of point-to-point (PTP) routing [26]. Rather than searching for the optimal path in a graph-based representation of the network, ACR features a seamless flow of the packet from source to destination based on a many-to-many communications paradigm [30, 31]. Any node receiving a packet will inspect its attributes based on which it decides whether to forward the packet or not. As such, the terms "routing" and "relaying" are used interchangeably throughout this chapter.

The MANET application scenarios considered herewith feature traffic flows which are predominantly convergecast. In other words, packets are unicast in the upstream direction to a single sink. To that end, current ACR schemes are not *fully* autonomous when it comes to unicast traffic. This is true since an end-to-end handshake must take place between each traffic source and the network sink. Such a handshake is necessary to define a "barrage" region (also referred to as a "suppression" region) between each source–sink pair. The said region serves to confine the traffic flow within certain geographical boundaries [28].

The handshake process required for spatial containment of traffic flows has to be revisited whenever significant topological changes occur, e.g., due to mobility [28]. To circumvent such a shortcoming, a novel method for constructing a fully autonomous cooperative routing (FACR) scheme is presented in this chapter. The method relies on the use of a novel physical layer (PHY) frame structure coupled with geographical (position-based) routing criteria.

Recognizing the advantages of ACR/FACR in addressing mission-critical application needs, this chapter unveils a few design challenges associated with these systems. The chapter mainly focuses on those prime challenges which are essential for any practical and technically feasible implementation of ACR/FACR. Practical hardware and software solutions to those challenges are presented and discussed in depth. The practicality of the proposed solutions is validated on software-defined radio (SDR) platforms.

The developed hardware and software is used to carry out field tests for the sake of empirical assessment of the performance, primarily the PHY layer. The end goal is to not only to offer a public-domain insight into how ACR/FACR can be practically implemented, but also on the outstanding throughput and latency performance of this class of MANET routing.

3 Autonomous Cooperative Networking Solutions

This section mainly aims to lay down the foundation for the subsequent discussion on the virtues of autonomous cooperative networking, and in particular how routing takes place. An analytical overview is provided with regard to why this relatively new class of MANET technologies not only challenges a stagnant MANET R&D ecosystem but also is better positioned to meet the aspirations of team-based mission-critical systems.

3.1 Autonomous Cooperative Routing Background

The goal of this subsection is not to offer a detailed literature survey of ACR-driven routing schemes. Rather, it aims at offering a brief historical background and some insight into the motivations for ACR. In the next subsections, some of the most prominent incentives for adopting ACR schemes are presented in more depth.

The field of mobile ad hoc networking (MANET) is a long-established field with a broad coverage in research literature. One of the most important topics addressed in MANET research is routing. For a long period of time, point-to-point (PTP) routing schemes (also known as path-oriented schemes [32]) were prevailing in literature as well as practical implementations [26]. Within this realm, geographical routing (geo-routing) has been widely accepted for routing in MANETs. This is mainly due to its resilience to mobility and network topological changes [33]. In fact, geo-routing was adopted by the European Telecommunications Standards Institute (ETSI) as a standard MANET routing scheme for Intelligent Transport Systems (ITS) [34].

Notwithstanding early signs of success, current implementations of geo-routing are highly likely to be plagued by an overhead that grows rapidly with node density and/or frame arrival rate [29]. This indeed has a negative impact on latency and throughput. Such an issue has already been identified as a priority to be addressed for scalable MANETs [23, 24].

Generally speaking, classical geo-routing schemes belong to one of two groups. The first is beacon-based whereby position beacons are exchanged between neighboring nodes, so as to maintain up-to-date topological awareness. On the other hand, beaconless geo-routing entails receiver-based contention to select the best packet forwarder [31]. Nonetheless, both forms suffer from the aforementioned problem: they are highly inclined to produce large overheads. This is either due to the repetitive

exchange of neighbor discovery messages (true for beacon-based protocols) or due to contention resolution overhead (applies to the beaconless case) [35].

Needles to reiterate that the routing protocol overhead must be decreased in order to meet the aspirations set forth for mission-critical applications. To serve exactly that purpose, cooperative transmission comes to rescue. From a conceptual point of view, autonomous cooperative relaying was first introduced in [36, 37]. The forwarding mechanism there was labeled as "randomized distributed cooperative transmission". Autonomous cooperative transmission was analyzed from the perspective of achievable cooperative transmit diversity in great depth in [38, 39].

In essence, autonomous cooperative relaying entails the forwarding of physical frames while not reverting to any relay selection process. The term *autonomous* mainly stems from the fact that nodes within a cooperative cluster are actually unaware of each other [26]. In other words, there does not exist any sort of cross-coordination between nodes before the frame is relayed. Therefore, autonomous cooperative routing is also often referred to as "blind cooperative transmission" [40].

The transformation of ACR concepts into practice entailed the need to find means for confining the packet flows spatially. Otherwise, unicast flows will quickly flood the network and unnecessarily hijack the spatial and temporal resources of the network [31]. As such, controlled barrage or suppression regions must be created by means of request-to-send (RTS)/clear-to-send (CTS) handshake between any arbitrary pair of source–sink nodes [26]. Traffic from a source to a given sink is suppressed and barred to spill outside the designated barrage region [30]. This line of work has been holistically treated in a series of chapters in [10, 25, 26, 28, 41]. To guarantee positive progress toward the sink, hop count to reach the sink is adopted as a routing metric.

In case position information is available to nodes, then position-based routing criteria can be used to streamline the forwarding process within narrow geographical corridors [29]. One possible manifestation of such approach is illustrated in Fig. 2. As shown in the figure, only nodes offering positive progress toward the sink take on the responsibility of forwarding the packet. Once a node receives a packet, it inspects the position attributes of the transmitters and compares them to its own. The PHY header has to be designed in a way that supports such a functionality as further described in Sect. 4.2. A simple geo-routing criterion is to for the receiver to forward the packet if it is closer than at least a certain number of transmitters.

The concurrent transmission of the same PHY frame provides for an array gain that is proportional to the number of transmitters at a given hop [30]. Such a gain contributes to the increase in the average hop distance and consequently reduces the e2e latency. Nonetheless, it also means substantially higher energy consumption per frame at a single hop. This important trade-off is analyzed and treated rigorously from an e2e perspective in [29]. It is shown there that for a given e2e energy consumption target, ACR can be tweaked to offer tangibly lower e2e latency.

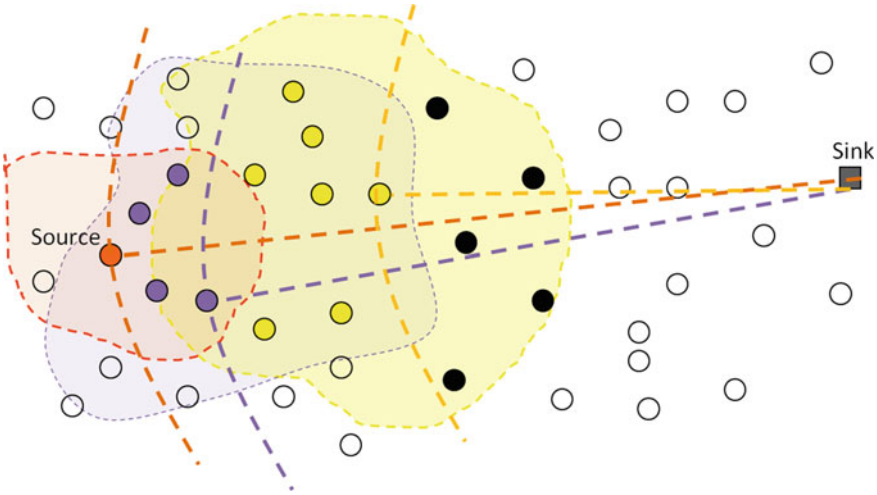


Fig. 2 Illustration of the operation of an autonomous cooperative scheme. A source injects a frame into the network. Receivers who are closer to the sink than the source will relay the frame. In the second hop, each receiver reads the position information conveyed by the transmitters of the first hop in order to decide whether to forward the frame or not. Any second-hop receiver offering positive progress toward the sink will forward it. The forwarding process continues seamlessly until the frame reaches its destination

3.2 Why Autonomous Cooperative Routing?

In this section, the advantages of ACR in comparison to path-oriented (i.e., PTP-based) routing schemes are studied. Three prime metrics are considered herewith: end-to-end latency, normalized transport rate, and maximum achievable throughput. A list of all notations used in the chapter is given in Table 1.

3.2.1 Lower End-to-End Latency

The end-to-end latency is given by $\sum_{q=1}^{Q} T_{h_q}$. Here, Q denotes the expected number of hops from source to sink assuming that a barrage (i.e., suppression) region has already been allocated for a given traffic flow. Further, T_{h_q} is the duration of the qth hop. In the case of ACR, the hop duration T_{h_q} is deterministic since there is no contention among potential relays. Accordingly, $T_{h_q} = T_p + T_t$, where T_p is the packet duration and T_t is the turnaround time corresponding to the change from receiver state to transmitter state.

On the other hand, path-oriented routing schemes entail some overhead pertaining to the selection of the optimal path. In the sequel, classical beaconless geo-routing is considered as a representative example. The sender needs to select the best relay, typically the one offering largest positive progress toward the destination. The selection

Table 1 List of notations

T_p	Data packet duration	T_t	Time to transition between Tx and Rx states
T_c	Control packet duration	T_w	Waiting time before successful channel access
T_b	Fixed back-off timer	T_{h_q}	Duration of the qth hop
Q	Average number of hops to destination	d	Hop distance
T_R	PHY header duration	P_t	Transmit power
P_n	Noise power	τ	Threshold on outage probability
γ_t	SINR for successfully decoding a frame	α	Large-scale path loss coefficient
γ_o	Mean SINR	a	Hop distance gain of ACR over PTP
I	Number of cooperative transmitters	N	Number of nodes in the network
$p(N)$	Probability that a node is a destination	$A(N)$	Area of the communications footprint
R_{max}	Maximum achievable per-node throughput	W	Over-the-air bit rate
P_b	Back-off probability	D	Source–destination separation
l_A	Packet arrival rate	l_e	Effective packet arrival rate in case of path-oriented routing
λ	Wavelength		
ρ	Node density	$\overline{M}$	Average number of nodes backing off
Δ_f	Subcarrier bandwidth	N_s	FFT size
$T_s = \frac{1}{N_s \Delta_f}$	Sampling time	$t = nT_s$	Discrete time representation
$k = -\frac{N}{2} \cdots \frac{N}{2}$	Subcarrier index	$i = 1 \ldots I$	Coop.transmitters
δ_i	CFO of ith Tx with respect to Rx	$a_{k,n} e^{j\phi_{k,n}}$	QAM symbol time n and subcarrier k
$h_{i,m}, m = 1 \ldots \mathcal{M}$	Channel fading coefficients	T	Time between multipath channel components
$T_i^{'}$	Propagation delay plus cooperative time offset associated with transmitter i	T_s	Signal sampling interval
		x_i^n	Phase rotation of the ith coop.transmitter
$N^{'}$	Number of nodes (barrage region update)	T_U	Total duration to update barrage regions
S_R	Number of times barrage update is repeated	$\mathcal{C}(\mathcal{UV})$	Number of hops measured from node $\mathcal{U}$ to V
B_L	Number of localization resource blocks	B_q	Number of relative position quantization tones
p_{st}	Probability of successful triangulation	F	Number of trials before successful triangulation
L	Number of bits in a data packet	P_o	Outage probability

is established through a handshake process. At the bare minimum, such a process consists of three transactions at *each* hop [35]:

(1) Request-to-send (RTS) message of duration T_c followed by a turnaround time T_t.
(2) A clear-to-send (CTS) message from the optimal receiver with duration that is also equal to T_c followed by T_t.
(3) Packet transmission with duration T_p.

The hop duration is therefore $T_{h_q} \geq T_p + 2T_t + 2T_c$ which is obviously always greater than that of the autonomous case. Therefore, end-to-end latency as a performance metric speaks in favor of ACR.

There is another factor that further boosts the latency performance of ACR. That is related to the fact that ACR exploits cooperative transmission techniques which in return feature array (power) gains [30]. Moreover, with some precoding and transmit-side signal processing, transmit diversity gains can be also attained [38]. By means of applying carefully selected randomization matrices to the transmitter vectors, such diversity gains can be obtained. To some extent, this approach has similar effect to the so-called phase dithering [42].

The array and diversity gains result in extending the average communication range compared to path-oriented routing. Undoubtedly, this results in reducing the number of hops Q, thus further contributing to the reduction of the end-to-end latency [43]. This is true since $\sum_{q=1}^{Q} T_{h_q}$ is a decreasing monotone in Q. A rigorous analysis of the progress made per hop in ACR networks can be found in [29].

3.2.2 Higher Normalized Per-Hop Transport Rate

The normalized transport rate (NTR) is defined as the average number of bits that can be communicated at a given hop over distance per unit time using one unit of energy [44]. The consideration of the normalized transport rate as a performance metrics stems from its ability to capture hop distance (which eventually affects end-to-end delay) as well as energy consumption.

Recalling that a contention phase ought to take place before a packet is routed in path-oriented schemes, then the upper bound on NTR is dictated by two factors:

(1) The minimum duration of a contention phase.
(2) The maximum achievable hop distance.

The hop distance has many definitions in literature, but here it is assumed that it refers to the positive progress made at a given hop along the line connecting the source to the destination. In the case of path-oriented schemes, assuming the mean hop distance to be equal over all hops is an acceptable approximation (especially in dense scenarios) [32, 45].

The hop distance, denoted by d, is governed by the underlying outage model. In a Rayleigh fading channel, the mean signal-to-noise and interference ratio (SINR) is given by

$$\gamma_o = \frac{2P_t}{P_n}\left(\frac{\lambda}{4\pi d}\right)^{\alpha}, \tag{1}$$

where λ is the wavelength, α is the large-scale path loss coefficient, P_t is the transmit power, and P_n is the noise power. The outage probability is given by $\mathrm{P}_o = 1 - e^{-\gamma_t/\gamma_o}$, where γ_t is the below which the receiver will be in outage. Consequently, the hop distance given a single transmitter d is expressed as

$$d \leq \left(\frac{\lambda}{4\pi}\right)\sqrt[\alpha]{\frac{2P_t \ln\frac{1}{1-\tau}}{P_n\gamma_t}}. \tag{2}$$

On the other hand, the duration of one complete contention phase cannot be shorter than one RTS message from the sender, one CTS message from the relay, plus the packet duration, T_p. As mentioned earlier, the half-duplex nature of the devices entails a turnaround time of T_t. Accordingly, the NTR for PTP-based routing is upper bounded by

$$NTR_{PTP} = \frac{Ld}{\left[2(T_c + T_t) + T_p\right]P_t(T_p + 2T_c)}, \tag{3}$$

where L is the length of a packet in bits. On the flip side of the coin, the NTR for the autonomous case is given by

$$NTR_{ACR} = \frac{Lda}{(T_p + T_R)^2 IP_t}, \tag{4}$$

where I is the number of cooperative transmitters, T_R is the duration of the PHY header, and a is a gain factor which reflects the fact that the hop distance in cooperative transmission mode is generally larger than PTP mode. There are indeed many factors affecting the value of a. Nonetheless, for the sake of simplification and conciseness of the analysis, the special case of I equidistant transmitters can be considered here. In such a case, $a = I^{1/\alpha}$. Accordingly, it can be shown from (3) and (4) that ACR-based systems offer better NTR under the condition that

$$I^{1-\frac{1}{\alpha}} < \frac{[2(T_c + T_t) + T_p](T_p + 2T_c)}{(T_p + T_R)^2}. \tag{5}$$

The values of T_p, T_c, T_t and T_R are mainly dictated by the underlying video transmission quality of service (QoS) requirements as well as hardware constraints.

In Sect. 6, a proprietary PHY implementation developed for this project is described in more detail. The implemented PHY is based on the use of orthogonal frequency division multiplexing (OFDM). The duration of one OFDM symbol is set at 8 μs. The duration of the PHY header is equal to 1 OFDM symbol. The preamble training sequence has the duration of exactly 38.4 μs. Therefore, the

shortest frame (i.e., one that is sufficiently large to carry an RTS or CTS control message) is $T_c = 54.4\,\mu s$. The turnaround time, T_t, is highly dependent upon the underlying radio front-end. In this specific implementation, it was measured to be around 180 μs.[2] Finally, the payload portion was set to consist of 50 symbols. While lower frames may be preferable from a frame error rate (FER) viewpoint, they are associated with larger PHY overhead ratio. A frame of 50 symbols, i.e., $T_p = 476.8\,\mu s$, strikes the right balance.

Based on the above, and assuming a path loss coefficient of $\alpha = 3$, then ACR outperforms PTP-based path-oriented scheme for $I < 3.02$. In other words, autonomous geo-routing offers higher NTR as long as is carried out by one, two, or three transmitters at a given hop.

3.2.3 Higher Maximum Achievable Throughput

The end-to-end latency performance is indicative but not sufficient to establish with evidence the superiority of ACR. Interference caused by other concurrent packet flows indeed has an adverse effect on e2e latency since it causes transmission outages and invokes back-off procedures. Hence, it must be taken into consideration. The interplay between interference and medium access is best captured by studying the maximum achievable throughput per node.

It was shown in [32] that ACR-based networks offer a per-node unicast capacity which scales in the order of $\Theta(\sqrt{N/\log N})$. This is identical to the Gupta–Kumar per-user capacity [46] that traditional path-oriented routing networks can offer. While such a result is reassuring, asymptotic scaling orders do not suffice to benchmark ACR against path-oriented PTP-based routing schemes. Furthermore, video streaming traffic in a mission-critical MANET is predominantly convergecast. As such, this must be taken into consideration.

Bisnik and Abouzeid provided a detailed throughput and delay analysis in a random access multihop network [47, 48]. For a network of N nodes, an absorption probability $p(N)$ is defined therewith as the probability that a traffic flow is terminated at an arbitrarily chosen node. It is straightforward to state that $p(N) = 1/N$ in a convergecast network.

Assuming a persistent back-off scheme [49], the mean waiting time before successful channel access is denoted by T_w. The back-off footprint, $A(N)$ is defined as the area around a given transmitter within which no other transmission can take place due to interference. $A(N)$ is actually normalized by the total area of the network. Finally, the maximum achievable throughput per node, R_{max}, is defined to be the maximum node throughput for which the end-to-end delay remains finite. Subsequently, R_{max} (in bps) is computed using [47], Eq. (22):

[2] A video capture of the turnaround time measurement is posted online for the interested reader (https://youtu.be/lDYVHZ6GcIM).

$$R_{max}(N) = \frac{Lp(N)}{T_w + \frac{L}{W} + 4NA(N)\frac{L}{W}}, \\ = \frac{L}{NT_w + NT_h + 4N^2A(N)T_h}, \tag{6}$$

where W is the bit rate.

The mean waiting time T_w is function of the back-off probability. The latter can be expressed as $P_b = \overline{M}/N$, where $\overline{M}$ is the average number of nodes that are forced to queue at least one frame of their own during the entire multihop journey of another frame [50]. Assuming Bernoulli distribution, the mean number of transmission attempts before success is $1/P_b$. As such, the mean back-off time can then be expressed as

$$T_w = (1 - P_b)T_b = \left(\frac{N}{N - \overline{M}}\right) T_b, \tag{7}$$

where $T_b \geq T_h$ is a fixed duration a node must wait before reattempting to retransmit.

To compute $\overline{M}$, first the probability that exactly m nodes will back off during a given hop is analyzed. Given n nodes exist in the back-off region and a packet arrival rate of l_A, then this probability is given by

$$p_m(m|n) = \binom{n}{m} \left(1 - e^{-T_h l_A}\right)^m \left(e^{-T_h l_A}\right)^{n-m}, m \leq n. \tag{8}$$

The probability that exactly n nodes actually exist in the region is

$$p_n(n) = \frac{1}{n!}(\rho A)^n e^{-\rho A}, \tag{9}$$

where ρ is the network node density under the assumption of 2D Poisson point process node distribution. Consequently, the probability distribution function of m is given by

$$p_m(m) = \sum_{n=m}^{\infty} p_m(m|n)p_n(n). \tag{10}$$

The next question to tackle: in light of the above, what is the probability, $P_M(M)$, that M sensor nodes backlog at least one transmission during the Q-hop lifetime of the packet in concern? The different permutations for distributing those M nodes over Q hops can be conveniently computed using integer set partitioning algorithms. These permutations can be expressed in matrix format as

$$\begin{bmatrix} m(1,1) & \dots & m(1,Q) \\ \vdots & \ddots & \vdots \\ m(\mathcal{P},1) & \dots & m(\mathcal{P},Q) \end{bmatrix} \in \mathbb{Z}^{\mathcal{P}\times Q}, \tag{11}$$

where $\mathcal{P}$ equals the number of different permutations corresponding to the distribution of M back-off nodes over Q hops. Consequently, the probability density function is obtained as follows:

$$P_M(M) = \sum_{u=1}^{\mathcal{P}} \prod_{i=1}^{Q} p_m(m(u,i)). \tag{12}$$

Therefore, a compact expression for $\overline{M}$ can be obtained as follows:

$$\overline{M} = \sum_{M=0}^{\infty} \sum_{u=1}^{\mathcal{P}} \prod_{q=1}^{Q} \sum_{n=m(u,q)}^{\infty} M p_m(m|n) p_n(n). \tag{13}$$

Substituting (13) into (7) gives the mean waiting time before successful channel access, T_w. It is paramount however to note that listen-before-talk (and consequently back-off procedures) is applied only once at the source in case of autonomous routing. On the other hand, it is applied at *each* intermediate hop in case of PTP-based (i.e., path-oriented) routing. In other words, the back-off procedures are invoked every time a node has a packet to send whether its own or an ingress packet from a neighboring node. Hence, the effective packet arrival rate in case of path-oriented routing is actually

$$l_e = \frac{l_A}{p(N)} = N l_A. \tag{14}$$

The computation of T_w and subsequently R_{max} is highly dependent on the mean number of hops, Q, as can be inferred from the analysis above. For a source–destination separation of D, the average number of hops in PTP-based systems is more or less $Q = \lceil D\sqrt{\frac{\pi}{A(N)}} \rceil$. On the other hand, such an approximation does not hold true in ACR-based systems. This is because in the long-term average sense, the hop distance grows in size every hop [29, 30]. As such, it is mandatory to derive a means to compute the probability mass function (PMF) of Q, which is the task to tackle next.

The probability of hopping Q times before reaching the destination is expressed as $p_Q(Q) = \mathbb{P}[x_Q \geq D]$, where x_Q is the total progress made after Q hops along the axis connecting the source and the destination. The number of cooperative transmitters at hop i is denoted by I_i. Further, the cumulative number of cooperative transmitters from the first hop till the $(Q-1)$th hop is given by $S_{Q-1} = \sum_{i=1}^{Q-1} I_i$.

An expression for the total progress made by the packet after Q hops was derived in Eq. (9) of [29] and is recalled here for convenience:

$$x_Q = \varphi S_{Q-1} + (Q-1)\frac{\beta}{U^{\frac{1}{\alpha}}} + x_1. \tag{15}$$

In (15), φ and β are network-dependent parameters, α is the large-scale path loss exponent, x_1 is the progress made in the first hop, and U is and outage-dependent constant that is given by [29]:

$$U = \frac{P_n}{2P_t}\left(\frac{4\pi}{\lambda}\right)^{\alpha}\frac{\gamma_t}{\ln\frac{1}{1-\tau}}. \tag{16}$$

It was also demonstrated in [29] that the PMF of S_{Q-1} can be computed for a given set of network parameters by recursion. Therefore, the PMF $p_Q(Q)$ can be computed numerically using

$$p_Q(Q) = \mathbb{P}\left[S_{Q-1} \geq \frac{1}{\varphi}\left(D - \frac{\beta}{U^{1/\alpha}}(Q-1) - x_1\right)\right]. \tag{17}$$

With the PMF readily available, the mean value of Q can be then easily computed.

The ratio of R_{max} for ACR to that of PTP-based was computed using (6)–(17). Results are shown in Fig. 3 in terms of the communication range gain, a. For a better and more insightful perspective, the e2e latency reduction factor that ACR enjoys over PTP-based routing is also plotted on the same figure. The plot in Fig. 3 is divided into 3 segments corresponding to the number of cooperative transmitters covering a given range of gain. Empirical results obtained from field testing and reported in Sect. 5 have been used to deduce the value of I (the number of cooperative transmitters) corresponding to a range of values for a (the ACR hop distance gain).

Although a larger gain favors ACR in terms of end-to-end latency, it is not always preferable in terms of throughput performance. It is evident from the figure that ACR starts to lose its edge in terms of per-node throughput as the gain increases. This is mainly because the coverage footprint of a packet transmission grows, thus blocking other nodes from accessing the network [32]. As such, it is essential to tune down the individual transmit power so that the gain is maintained within limits.

3.2.4 Summary

It is worthwhile at this point to summarize the key findings so far. To compare ACR to classical path-oriented routing schemes, it is best to fix the NTR as a performance constraint since it is the one that captures energy consumption. It has been already shown that with $I = 3$, ACR and path-oriented schemes offer the same NTR. However, it is clear from Fig. 3 that ACR offers up to 2X improvement in the maximum achievable throughput per node. It can be also inferred that ACR enjoys at least 4X reduction in the end-to-end latency. A corollary to this statement is that if throughput and latency targets are fixed, ACR will consume substantially less energy per transported bit. Looking at it from either perspective, ACR outperforms classical PTP-based path-orient MANET routing by far.

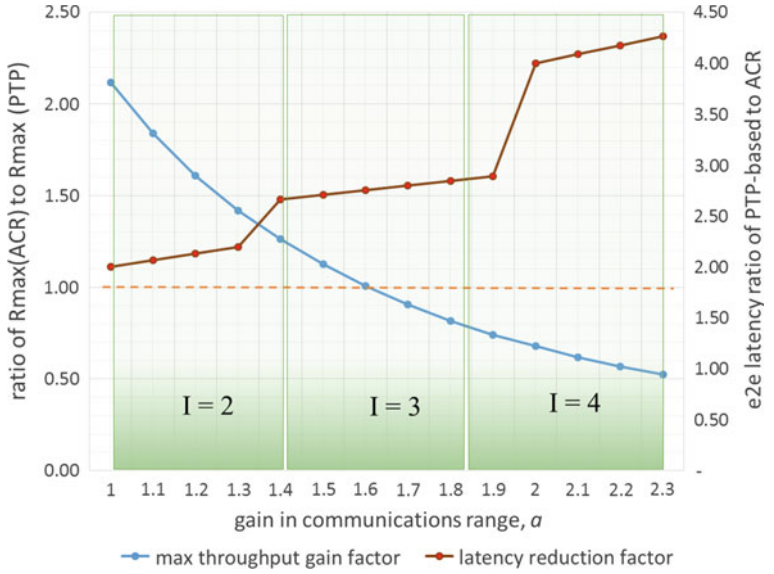


Fig. 3 The performance of ACR is compared to PTP-based routing in terms of maximum achievable per-node throughput as well as end-to-end latency. Using the analytical results of Sect. 3.2, the computations were carried out assuming a network of 20 nodes at a density of $\rho = \frac{1}{30^2}\text{m}^{-2}$. The average PTP communication range was $\leq$20 m at a path loss coefficient of 2.8. Packet duration of 0.5 ms is assumed at an arrival rate of 100 s^{-1}

A final note should be tailored for security aspects. Indeed, security is a paramount concern in mission-critical applications and should not be overlooked. However, it is quite an involved task to benchmark the performance of ACR protocols to path-oriented ones in terms of susceptibility to malicious attacks. Recognizing its importance, analysis of security aspects for ACR is left as future scope of work.

4 A Higher Degree of Autonomy

While autonomous cooperative routing (ACR) has been shown to offer an undeniable value proposition, there is more that can be achieved within its realm. In this section, we describe the motivation for developing a *fully* autonomous cooperative routing technique. We also highlight some of the key design elements as well as practical implementation considerations.

4.1 Motivation

Route stability is defined as the probability that an end-to-end path connecting source to destination is still available after a certain duration from being established [51]. Indeed, ACR has been shown to offer better route stability compared to path-oriented MANET routing schemes under realistic mobility models [28]. In other words, ACR-based MANETs are substantially more tolerant to topological changes. Nevertheless, barrage regions still need to be maintained and regularly updated.

A barrage region must be initially created then regularly updated for each source–destination pair. In a convergecast mode (which is typical in mission-critical applications), this mandates the execution of a round-trip end-to-end handshake between each node in the network and the network sink. For a given node $\mathcal{U}$, the handshake process between $\mathcal{U}$ and the sink is essentially designed so that all other nodes can measure how many hops away from $\mathcal{U}$ they are [25]. This is then used to carve the barrage region from $\mathcal{U}$ to the sink. Denoting the hop count from $\mathcal{U}$ to another node $\mathcal{V}$ with $\mathcal{C}(\mathcal{U}\mathcal{V})$, then a simple rule is to have nodes with $\mathcal{C}(\mathcal{U}\mathcal{V}) > Q(\mathcal{U})$ suppress the transmission of $\mathcal{U}$'s packets [26].

While traffic in convergecast networks is predominantly upstream (traffic toward the sink), there is the need to cater for downstream traffic encompassing control and configuration messages. It is inaccurate however to consider that the barrage region in the upstream direction is good enough to represent that in the reverse direction, i.e., downstream. Reciprocity on weighted graphs (such as wireless networks) is a highly contentious issue [52]. As such, routes are generally nonreciprocal, and consequently, a barrage region has to be separately created for each direction of the traffic.

There are surely multiple approaches to manage the barrage region update process. It is important to note that the network can handle only one handshake process at a time. This is true since messages emanate from or terminate at a single node. Therefore, it is quite challenging to handle more than one update process at a time due to interference constraints. In other words, the network sink is required to orchestrate the barrage region update process. Assuming the sink has prior knowledge of all mobile nodes in the network, then one feasible approach consists of three phases:

(1) A broadcast message from the sink soliciting a response from node $\mathcal{U}$. Intermediate nodes relaying the response message increment a designated hop count field in the packet as it traverses the network toward $\mathcal{U}$.
(2) A response message which is broadcast from node $\mathcal{U}$ back to the sink. Any intermediate node relaying the response message performs two tasks:
    (a) Increments a designated hop count field in the packet as it traverses the network toward the sink.
    (b) Takes a decision whether it lies within or outside the downstream barrage region of $\mathcal{U}$.

(3) To shape the upstream barrage region of $\mathcal{U}$, the sink has to rebroadcast another message containing the hop count $Q(\mathcal{U})$ measured on the previous message. As this message traverses the network, each intermediate relay node decides whether it belongs within or beyond the upstream barrage region.

The process above is then sequentially repeated across the whole node population. Putting things into perspective, as the number of nodes N gets larger, the barrage region update process starts to have a tangibly significant overhead. This issue is discussed next.

In mission-critical operations, it is reasonable to mandate that all of the nodes complete the barrage creation/update process. Otherwise, nodes which are left out (for one reason or another) will resort to broadcasting, i.e., flooding, all of their frames. Undoubtedly, this causes substantial interference and unnecessarily overgrazes the network's spatial and temporal resources. As such, the barrage region creation/update process should target a 100% reachability. Reachability is a metric that measures the percentage of nodes which can be covered, i.e., are reachable, after performing $\mathcal{X}$ broadcast rounds [53]. Reachability is denoted by a positive monotonic function $R(\mathcal{X}) \leq 1$, where $\mathcal{X} = 0 \ldots \mathcal{X}_{max}$, $R(\mathcal{X}_{max}) = 1$, and $R(0) = 0$.

The barrage region handshake process has to be effectively executed with each node as many times as needed to reach that node. This actually contributes to increasing the duration of the barrage creation/update process. Subsequently, the effective number of nodes can be essentially defined as the number of times the handshake process is executed until barrage regions for all nodes have been established. Taking into consideration the fact that the handshake process consists of three broadcast phases, then the effective number of nodes is therefore given by

$$N^{'} = N \left(\sum_{\mathcal{X}=1}^{\mathcal{X}_{max}} \mathcal{X} \left[R(\mathcal{X}) - R(\mathcal{X}-1) \right] \right)^{3}. \tag{18}$$

Another major factor to be considered relates to the fact that the hop count is not a deterministic parameter but rather a discrete random variable. This is a fact of crucial importance since the hop count from the source to the sink as well as to the intermediate nodes is the sole parameter used in defining the barrage region [26]. The number of hops measured from the traffic source $\mathcal{U}$ to an intermediate relay node $\mathcal{V}$ at around $\mathcal{X}$ is denoted by $\mathcal{C}_{\mathcal{X}}(\mathcal{UV})$. As a matter of fact, (17) can be used to derive the PMF of $\mathcal{C}(\mathcal{U})$ by substituting Q with $\mathcal{C}_{\mathcal{X}}(\mathcal{UV})$ and making D equal to the distance between the $\mathcal{U}$ and $\mathcal{V}$.

Denoting the average hop count by $\overline{\mathcal{C}}(\mathcal{UV})$, it can be demonstrated numerically that the probability $\mathbb{P}[\mathcal{C}_{\mathcal{X}}(\mathcal{UV}) \neq \overline{\mathcal{C}}(\mathcal{UV}]$ has an appreciable value. An immediate conclusion can be drawn: the three-way handshake process must be carried out more than once for each node, i.e., $S_R \geq 2$ times, in order to come up with an acceptable estimate of $\mathcal{C}(\mathcal{UV})$. Analysis of S_R and its relation to the confidence intervals of $\mathcal{C}(\mathcal{UV})$ is actually left off as follow-up work to this chapter.

Based on all of the above, the total time required to finish the barrage region creation/update process is given by

$$T_U = 3Q_{max}N^{'}T_hS_R, \tag{19}$$

where Q_{max} is the maximum number of hops required for the broadcast message to reach all nodes in the network. It is insightful at this point to put things into perspective using a numerical example. In [28] (Fig. 4), it was shown that path availability probability drops below 95% after approximately 25–50 s.[3] Tactical and mission-critical MANETs can typically have as many 100 nodes [26]. Nodes are spread out such that up to 10 hops may be needed for a broadcast message to cover the network [28]. The hop duration can be assumed to be in the range of $T_h = 500\,\mu$s which includes a very short packet transmission time, processing time, and radio turnaround time. The effective number of nodes is highly influenced by $R(1)$ which is typically in the range of 95% [53]. Taking $R(\mathcal{X}) = [0.950, 0.990, 0.999, 1.000]$, then $N^{'} = 119$. Assuming $S_R = 2$, then (19) yields a whopping $T_U = 3.57$ s! This is at 14–28% contribution to the protocol overhead.

The barrage region creation/update overhead should also account for cases of network entry, i.e., new nodes joining the network. Join events will cut off the live network operation for a non-negligible period of time. So based on all of the above, there is sufficient rationale and motivation to fortify ACR with *full* autonomy, the subject of which is discussed in the next section.

4.2 Full Autonomy Enabled by Geo-routing

What would it take for a node to locally decide whether it should forward a given source's packet or not? What if a node is equipped with the capability to qualify whether its participation in the forwarding process is beneficiary to the packet's progress toward the sink? The availability of such a capability unleashes fully autonomous cooperative routing.

Knowledge of position relative to the sink is sufficient to meet that goal. During network initialization phase, the source sends a broadcast packet informing all other nodes of its position. Each node is also required to acquire its position relative to the sink. This can be done by means of an onboard global positioning system (GPS) module. Contrary to the classical perception, low-power GPS modules have been commercially available for quite some time. As a matter of fact, power consumption by the GPS module is far less significant than other key components in wireless

[3]The choice of a value for the path availability metric is indeed relative and subject to the underlying application. In mission-critical applications, robustness and high reliability are often stressed as key performance indicators by end users. Thus, selecting 95% as a benchmark mainly stems from feedback the authors accumulated through interactions with end users.

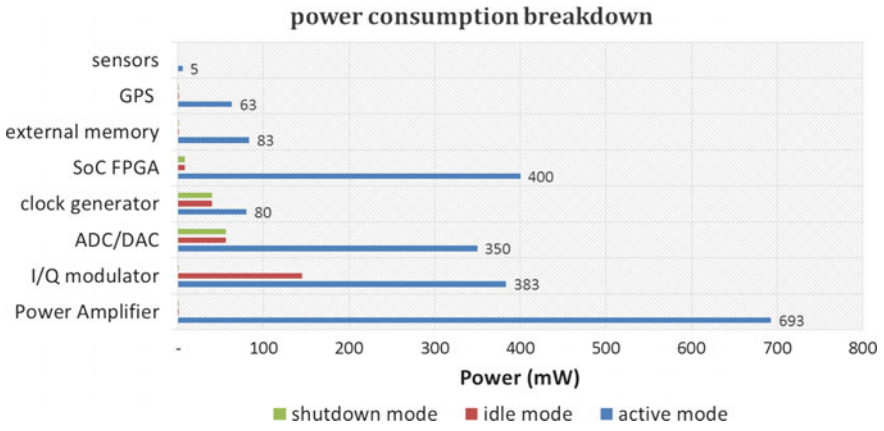


Fig. 4 Various candidates for each component in the system have been surveyed by the authors with low power consumption as a prime objective. The survey quickly revealed that the analog front-end components (i.e., I/Q, PA, ADC/DAC) are the most power-hungry. Duty-cycling these components whenever possible is not only a good practice but a necessity. The power budget of the GPS module can be considered as insignificant

communications systems. For instance, the analog front-end is far more power-hungry than the GPS module as illustrated in Fig. 4. Furthermore, the GPS module can be deeply duty-cycled to further save power.

The availability of position information allows the coupling of geo-routing and autonomous cooperative transmission. The result is full autonomy. This has been already eluded to in the illustration offered in Fig. 2 in Sect. 3.1. Full-length GPS positions are not really needed. Instead, each node needs to compute its relative position vector (distance and azimuth) with the origin being the sink. Furthermore, the system design has to cater for the very likely situation of weakening or complete blackout of the GPS signal.

Fortunately, the timescale of node mobility is quite relaxed: losing the GPS signal for a few seconds is likely to induce only intangible changes in the network topology. So it is more of an opportunistic approach which is advocated herewith where the position vector is updated whenever the GPS signal is accessible. Nonetheless, to account for those cases where a subset of nodes may suffer from prolonged GPS signal loss, a cooperative localization method is presented later in this section.

From a practical point of view, the challenge concerns the means by which cooperative transmitters can convey their position information to receivers (i.e., nodes which are the potential next-hop forwarders). An inherently related challenge is for this means to concurrently support the self-localization capability. The solution addressing both requirements is presented in the next subsection.

4.3 Random Access

To facilitate the communication of position information by transmitters, random access resources are allocated within the PHY frame [54] as shown in Fig. 5. The random access (RA) area consists of two distinct parts. The first one contains a total of B_Q tones which are allocated for progress quantization purposes. The second part is consists of B_L resource blocks distributed over b OFDM symbols and are allocated explicitly for localization purposes. The design and processing considerations of the localization part of the RA area is discussed in the next subsection.

Before a cooperative transmitter sends a frame, it quantizes the progress it offers with respect toward the sink. There are B_Q quantization levels such that resolution is D/B_Q, where D is the distance between the source and the sink. Each step is allocated exactly one tone in the random access area shown in Fig. 5. The relay needs to indicate the quantized progress it offers by simply energizing the corresponding tone whose index is equal to its progress level. Simple on-off keying (OOK) binary modulation is used to modulate the respective tone. At the receiver side, the B_Q tones will be routed from the output of the FFT stage toward the OOK demodulator as shown in Fig. 6. Progress levels of the respective transmitters are extracted and fed to a routing decision module.

Again, it is worthwhile to put things into perspective from a practical point of view. Nodes can be assumed to be distributed over a finite 2D disk with diameter D_{max} according to a binomial point process (BPP) [55]. However, in a geo-routing context, the progress along the line connecting source to destination is what really matters. As such, the 2D BPP distribution can be projected or more precisely reduced to a 1D

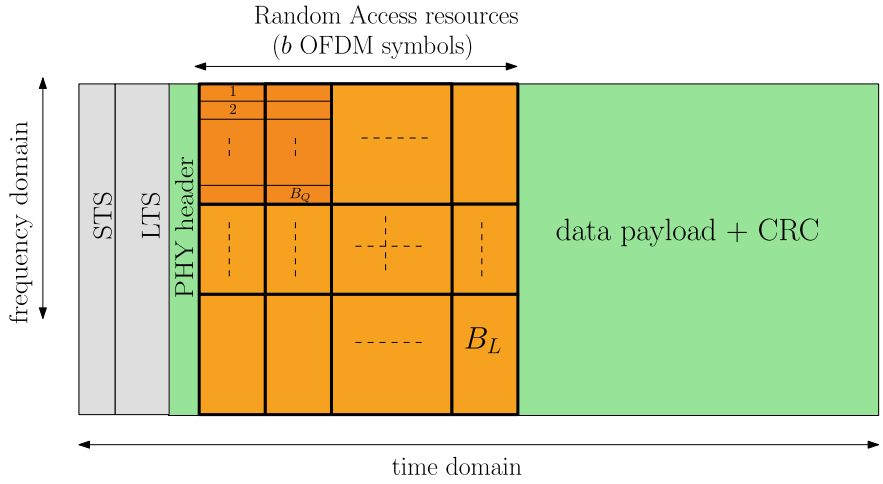


Fig. 5 A random access (RA) area is inserted into the OFDM frame to support two capabilities: (1) allow cooperative transmitters to indicate the progress they offer toward the sink, and (2) encode their position information that can be used by receiver so as to perform a TDOA-based self-localization

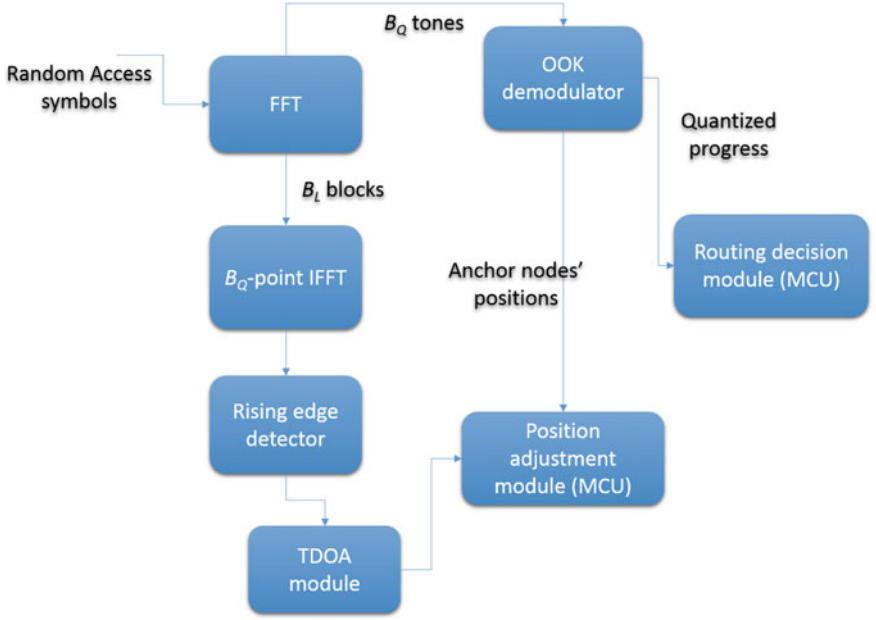


Fig. 6 A block diagram illustrating the processing of the quantization tones and localization resource blocks, both part of the random access area illustrated in Fig. 5

distribution. Consequently, the average distance to the ith nearest neighbor along the 1D progress dimension is given by $\frac{1}{2}ID_{max}/(N+1)$ [55]. The progress quantization level must be made sufficiently small to accommodate node displacement patterns. One viable design criterion is to have the quantization step D_{max}/B_Q larger than the distance to the Ith nearest cooperative transmitter along the progress line. In other words, it is to have

$$B_Q \geq 2(N+1)/I. \tag{20}$$

For $N = 100$ and $I = 3$ nodes, then $B_Q \geq 68$ tones which can be easily allocated within the stretch of one or two OFDM symbols.

4.4 Self-localization Scheme

It has already been shown by [28] that it takes 25–50 s before the end-to-end path starts to become obsolete under realistic mobility models. A corollary to this is that nodes can afford to lose their GPS signals for an equivalently long duration. Nonetheless, there might be situations where some nodes may suffer from GPS signal blackouts for even longer durations. Mission-critical systems have to incorporate higher levels

of resilience and robustness by definition and therefore need to account for such corner cases.

Nodes can capitalize on the presence of the random access area to carry out a triangulation procedure [54]. Those nodes which enjoy clear GPS signals can transmit their position information on regular basis so that others without GPS access localize themselves. As shown in Fig. 5, the random access area incorporates B_L resource blocks just for that purpose.

The method proposed for self-localization is to compute time difference of arrival (TDOA) [54]. Therefore, localization resource blocks need to cater two pieces of information: position information of the transmitters and propagation delay differences. The first one is straightforward and entails each anchor node encoding its position information into one of the localization blocks. A block is selected randomly by an anchor node and therefore collisions may occur. This is further discussed at the end of this subsection. Within this context, anchor nodes simply represent that subset of transmitter nodes which still enjoy clear access to the GPS signal.

On the other hand, extraction of TDOA information exploits the fact that each uniquely selected resource block contains a signal with a unique time signature. This is further illustrated in Fig. 7. The time reference at the receiver is influenced by the first energy arrival in the preamble portion of the frame. The B_L time waveforms must be reconstructed in order to detect the offset of each one from the zero time reference. As such, the B_L blocks are fed sequentially back to a B_Q-point IFFT module as depicted in Fig. 6. The TDOA can then be measured.

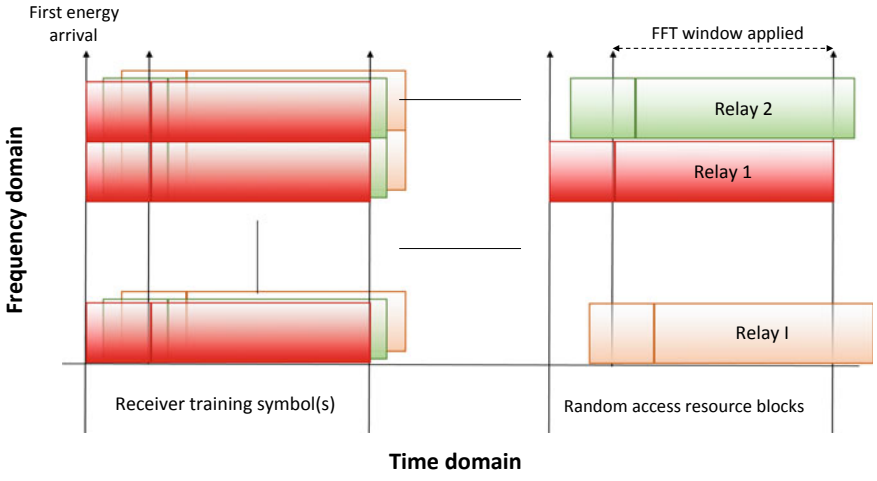


Fig. 7 The localization random access resource blocks are offset from each other in time. This is due to the fact that each block is modulated by a different transmitter (obviously as long as it happens to be selected by one transmitter)

Inherent to any random access methodology, collisions may occur. Therefore, a sufficient number of resource blocks B_L should be allocated. It has been shown in [29] that when I nodes randomly access B_L random access resource blocks, the probability of at least $z \leq B_L$ uniquely selected blocks can be evaluated recursively using

$$p_z = p_{z-1} \left(\frac{B - z}{B - z + 1} \right)^{I-z}, \tag{21}$$

where $p_o = 1$. For triangulation purposes, at least three nodes are required. Subsequently, the success probability of self-localization for given received frame is given by

$$p_{st} = \prod_{z=1}^{3} \left(\frac{B_L - z}{B_L - z + 1} \right)^{I-z}, B_L \geq 3. \tag{22}$$

The number of frames until the triangulation function succeeds is denoted by F. Had I been constant, the mean F would have been represented by a geometric random variable whose mean is $\mathbb{E}[F] = 1/p_{st}$. Nonetheless, I is also randomly distributed and understanding its statistical behavior is nothing but trivial. This is true especially since the value of I depends on a multitude of factors including packet forwarding statistics and GPS signal loss patterns.

Having said that, I is expected to grow whenever the receiver is closer to the network sink and/or the GPS signal is less likely to be blocked. If I tends to be large, $\mathbb{E}[F]$ will also be, i.e., it will take a few frames before a node with lost GPS signal can triangulate itself. Fortunately, however, when I tends to be large, this also implies that the expected number of nodes with lost GPS signal is small.

In any case, one can obtain a practical flavor of $\mathbb{E}[F]$ by noting that it is upper bounded by $1/p_{st}$ (evaluated at $\mathbb{E}[I]$). This is true by means of Jensen's inequality since it can be directly shown using (22) that $\mathbb{E}[F]$ is strictly concave in terms of I. The value of $\mathbb{E}[F]$ has been computed for a range of $\mathbb{E}[I]$ and results are reported in Table 2. The table clearly shows that with only $B_L = 5$ blocks, there is ample

Table 2 Average number of frames required until triangulation succeeds. A total of $B_L = 5$ resource blocks are assumed to be allocated in the random access area

$\mathbb{E}[I]$	Situation	$\mathbb{E}[F]$
3	Node closer to the network perimeter and/or heavy GPS signal blockage	2.08
5	Node in the middle of the network	13.02 and/or mild GPS signal loss
7	Node close to the network sink and/or low likelihood of GPS signal loss	81.38

time for nodes to adjust their positions. For the worst case scenario of $\mathbb{E}[I] = 7$, and assuming 1-ms frames, it takes no more than 82 ms to update the position.

5 Practical Implementation Challenges

The goal of this section will be tailored toward some of the practical challenges related to the implementation of ACR/FACR. Most of these challenges mainly stem from the nontraditional wireless channel characteristics in a cooperative transmission setup. As such, this section starts off with the presentation of the channel model which is cooperative by design. It then immediately delves into PHY design challenges invoked by the cooperative channel. Remedies and solutions are highlighted as well throughout the section.

5.1 Wireless Channel Model

From a PHY perspective, ACR in principle is a technique that allows multiple nodes to transmit the same frame almost concurrently. This statement needs to be further reinforced with respect to two different timescales. Concurrency is really true only at the packet level. At the symbol-level, however, the cooperative transmitters are not perfectly aligned in time and they need not be. In other words, the channel model has to accommodate the case of asynchronous transmission case.

In most recent literature, the case of asynchronous cooperative transmission has been referred to as the cooperative time offset (CTO) [56]. Even in the case of perfect synchronization among the I cooperative transmitters (e.g., by means of having access to GPS), there will still be time offsets from the receiver perspective due to propagation delay differences. Both effects are captured in the cooperative channel model by introducing the delays $T_1^{'} \ldots T_I^{'}$ as illustrated in Fig. 8.

The channel between an arbitrary pair of nodes is represented by a generic wide-band frequency-selective multipath tap-delay line with Rayleigh-distributed tap gains [57]. On average, there are $\mathcal{M}$ such taps. Natural echoes due to multipath are grouped in intervals of duration of T seconds. Mobility speeds are with the pedestrian to slow vehicular ranges such that the fading coefficients are assumed to be quasi-static, i.e., they remain constant during a single frame.

Orthogonal frequency division multiplexing (OFDM) is employed as a measure to counteract that frequency selectivity of the cooperative channel. The duration of the OFDM symbol is assumed to be larger than $(\mathcal{M}-1)T + \max\{T_i^{'}\}_{i=1}^{I} - \min\{T_i^{'}\}_{i=1}^{I}$ ensuring that each subcarrier encounters approximately a frequency-flat fading [58]. Amending each OFDM symbol with a cyclic prefix eliminates inter-carrier interference (ICI) and restores orthogonality between subcarriers. This enables decoupled signal detection at each subcarrier.

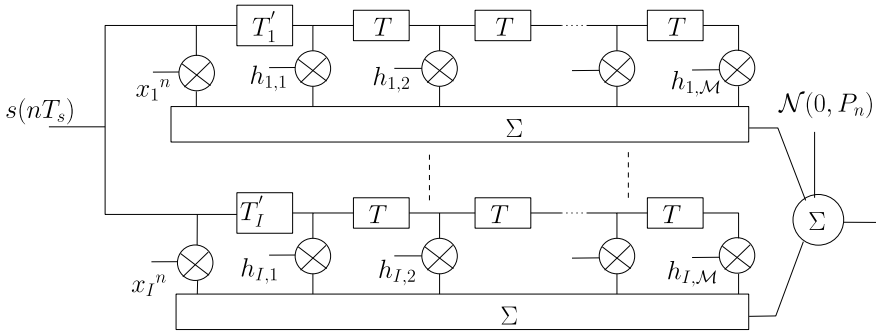


Fig. 8 Composite channel response capturing CFO plus Doppler spread, propagation delay differences, as well as multipath channel effects

Under the reasonable assumption that the fading coefficients $h_{i,m}$ are all mutually independent, it follows that $H(f)$ is complex Gaussian such that $H(f) \sim \mathcal{N}(0, \sigma_S^2)$, with

$$\sigma_S^2 = \sum_{i=1}^{I} \sum_{m=1}^{\mathcal{M}} \mathbb{E}[|h_{i,m}|^2]. \tag{23}$$

Furthermore, $|H(f)|^2$ is exponentially distributed with a mean of $2\sigma_S^2$. We note that $\sum_{m=1}^{\mathcal{M}} \mathbb{E}[|h_{i,m}|^2]$ represents the mean power content of the channel between the receiver and the ith transmitter and is equal to $(\lambda/4\pi d)^{\alpha}$. Therefore, we obtain

$$\sigma_S^2 = \left(\frac{\lambda}{4\pi}\right)^{\alpha} \sum_{i=1}^{I} \frac{1}{d_i^{\alpha}}. \tag{24}$$

It is assumed that the duration of the cyclic prefix of the OFDM symbol is long enough such that all signal echoes (natural and artificial) arrive within the cyclic prefix interval. Other ongoing packet relaying processes will rather contribute to the interference signal. This interference, however, will be also Gaussian since the individual channel gains are Gaussian [59]. The exact nature of such an external interference is beyond the scope of the present work.

5.2 Cooperative Carrier Frequency Offset

The sampling frequency is $1/T_s$, and n is a running sample index. The number of subcarriers is denoted by N_s. Due to clock imperfections, a carrier frequency offset (CFO) naturally exists between any arbitrary pair of nodes [43, 54]. The CFO between transmitter i and the receiver is denoted by $\delta_i^{(CFO)}$. The cooperative carrier frequency offset (CCFO) is defined herewith as $\max_i \delta_i^{(CFO)} - \min_i \delta_i^{(CFO)}$. On the other hand,

all nodes are assumed to be mobile; thus, a frequency Doppler exists between the ith transmitter and the receiver and is denoted by $\delta_i^{(DOP)}$. The CFO and the Doppler shift together have the combined effect of causing a phase rotation. Such an effect is captured in the model of Fig. 8 by defining:

$$x_i \triangleq e^{j2\pi\delta_i \frac{1}{N\Delta_f}} \ , \ \delta_i = \delta_i^{(CFO)} + \delta_i^{(DOP)}. \tag{25}$$

Hence, taking the individual CFO and Doppler shift effects into consideration, the baseband signal transmitted by node i is expressed as

$$\begin{aligned} s_i(nT_s) &= \sum_{k=-\frac{N_s}{2}}^{\frac{N_s}{2}} a_{k,n} e^{j\phi_{k,n}} e^{j2\pi(k\Delta f+\delta_i)nT_s} \\ &= x_i^n \sum_{k=-\frac{N_s}{2}}^{\frac{N_s}{2}} a_{k,n} e^{j\phi_{k,n}} e^{j2\pi kn/N_s}, \end{aligned} \tag{26}$$

where $a_{k,n}e^{j\phi_{k,n}}$ is the transmitted symbol. Consequently, the frequency-domain response of the composite channel is given by

$$H(n,f) = \sum_{i=1}^{I} {x_i}^n e^{-j2\pi f T_i^{'}} \sum_{m=1}^{\mathcal{M}} h_{i,m} e^{-j2\pi f(m-1)T}. \tag{27}$$

From (27), it is clear that the channel is highly time-varying because of the CCFO. This is true even though the fading coefficients are assumed to be quasi-static.

The time-varying nature of the channel mandates robust receiver design. As a matter of fact, the detrimental effect of the CCFO is far more adverse than that of the Doppler spread alone. This is true since the CCFO can be orders of magnitude larger. This is better appreciated by means of an example. A 1-ppm free-running clock yields a CFO around ± 2400 Hz at a center frequency of 2.4 GHz. In comparison, the maximum Doppler shift for a node moving at 10 km/hr, for example, is less than 25 Hz. Consequently, it is evident that the CCFO problem is order of magnitude more challenging than the classical Doppler spread problem.

In the presence of CCFO, the channel coherence time (roughly equal to the 0.423 times the reciprocal of the maximum Doppler shift [57]) in case of free-running clocks is comparable to the duration of just few OFDM symbols. The CCFO poses a couple of serious challenges on receiver design which has to cater for such a highly dynamic and fast-changing condition. Two of such challenges along with viable remedies are outlined in the following.

5.2.1 Automatic Gain Control Aging

The purpose of automatic gain control (AGC) in the receiver is to perform preamplifier gain adjustments. These adjustments are required in order for the signal to be received within the dynamic range of the analog-to-digital converter (ADC) [60]. The AGC module typically operates on the preamble portion in the very beginning of the PHY frame. It is in essence a feedback control loop whose goal is to maximize the input signal within the linear range of the ADC.

The correlation coefficient between two time samples of the Rayleigh fading envelope separated by τ_v is given by the zeroth-order first kind Bessel function $J_o(2\pi\delta_i\tau_v)$ [57]. For illustration purposes, the correlation coefficient for the case of zero CCFO (i.e., in the presence of only Doppler shifts) is compared to a 500-Hz CCFO on the same timescale in Fig. 9.

From Fig. 9, it is apparent that the AGC gain value will quickly become outdated in the presence of CCFO. This is also referred to as AGC aging. By the end of the frame, the outdated AGC value will be either:

(1) Too high, therefore driving the incoming signal to the nonlinear range of the ADC and causing significant signal distortion.
(2) Unnecessarily too low, thus the received signal may suffer from a severe SNR drop.

To address the AGC aging problem, there is the obvious option of using shorter PHY frame durations. Nevertheless, this will indeed increase the PHY overhead ratio and hence adversely affect the throughput. A more preferable option is to rerun the

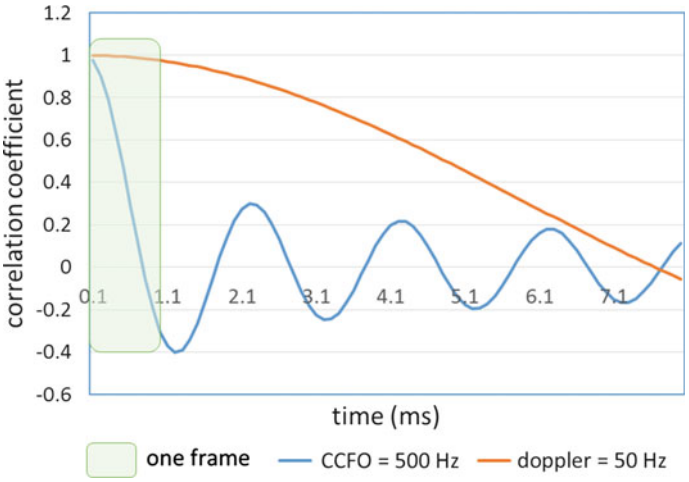


Fig. 9 Correlation coefficient of the fading channel envelope in case of zero and 500-Hz CCFO. In both cases, the Doppler shift is assumed to be 50 Hz. It is evident that CCFO produces a highly time-varying channel, and consequently, the channel gains quickly become uncorrelated even within the timescale of a single PHY frame

AGC module on pilots tones which are inserted within the PHY frame. The AGC loop may take quite a few samples in the beginning of the frame to converge. This is true since the channel variation from frame to frame may be unpredictably high. A new frame is a new transmission with a new set of cooperative transmitters. Hence, the power of the incoming signal is uncorrelated to that of the previous frame. On the other hand, the convergence time of the AGC loop when run on pilots is much faster. This is true since the channel variation within a frame is statistically correlated.

5.2.2 Aging of Channel Estimates

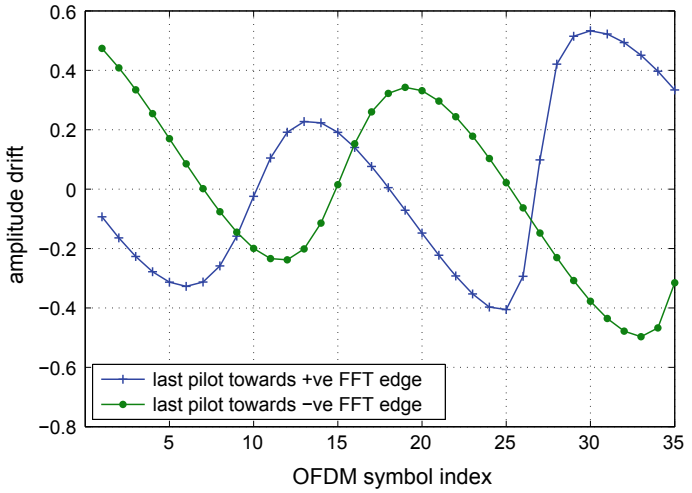
Similar to the AGC, channel estimation is typically performed on the preamble portion as well. Specifically, the preamble includes a long training sequence (LTS) symbol that is known a priori to the receiver. In OFDM systems, the LTS is used to estimate the fading channel coefficients corresponding to each frequency subcarrier [56]. Due to the highly time-varying nature of the channel, the estimates of the fading coefficients obtained in the beginning of the frame quickly become obsolete. In the presence of CCFO, the coherence time of the channel can be much shorter than the frame duration. As such, channel estimates need to be updated more often.

One approach is to insert more training symbols (i.e., LTS symbols) within the payload portion of the PHY frame. However, this will significantly increase the PHY overhead. This is true particularly since the coherence time is too small. For example, if the CCFO is 1000 Hz, then the coherence time of the channel is about 425 μs. A good design practice is to ensure up-to-date channel coefficients at least at a rate of 10 times per coherence window, i.e., an LTS symbol must be inserted once every 42.5 μs. For a symbol duration of 8 μs, this means that an LTS must be inserted at least after every 5th symbol. Hence, the overhead contribution of channel estimation is in excess of 16% which is quite significant.

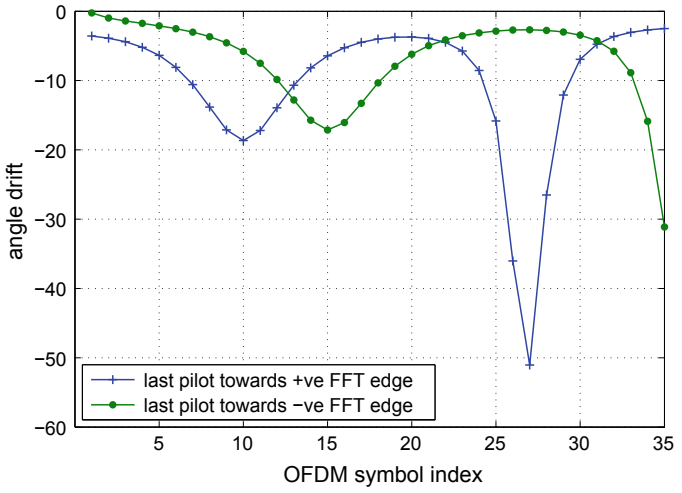
In an attempt to relax such an overhead, one may argue for farther spacing LTS symbols in the time domain. Such a proposition would entail the use of linear interpolation to compute the amplitude and phase of the channel coefficients for OFDM symbols in between the LTS symbols. However, as Fig. 10 strongly suggests, the level crossing nature of the cooperative channel is quite aggressive thus rendering the linear interpolation option very risky.

A neater approach, on the other hand, is to autonomously estimate the channel in a continuous fashion using the well-known decision-directed estimation (DDE) method [61]. Each OFDM symbol consists of N_s samples. At the end of the LTS (which is the first symbol in the frame), the least squares (LS) channel estimate at subcarrier k is given by

$$\widehat{H}(1,k) = \frac{\sum_{n=1}^{N_s} s^*(N_s - n, k) r(N_s - n, k)}{\sum_{n=1}^{N_s} |s(N_s - n, k)|^2}, \tag{28}$$



(a) Amplitude drift over time.



(b) Phase drift over time.

Fig. 10 Cooperative transmission in the presence of CCFO induces significant fluctuations in the phase and amplitude of channel fading coefficients. Thus, linear interpolation is by large infeasible

where $r(n, k)$ is the received signal observed at the nth time sample at the kth sub-carrier of the fast Fourier transform (FFT) stage output. In (28), $s(n, k) = p(n, k)$ when k is pilot tone, otherwise $s(n, k) = \hat{s}(n, k)$, i.e., the decided symbol. To obtain the channel estimate at any other arbitrary symbol $z = 2, 3, \ldots$, recursive estimation can used as follows:

$$\hat{H}(z,k) = \frac{\hat{y}(zN_s,k)}{\hat{Y}(zN_s,k)}, \tag{29}$$

where

$$\hat{y}(z+1,k) = \hat{y}(z,k) + r(z,k)s^*(z,k) - r(z-1,k)s^*(z-1,k) \tag{30}$$

$$\hat{Y}(z+1,k) = \hat{Y}(z,k) + |s(z,k)|^2 - |s(z-1,k)|^2. \tag{31}$$

A DDE receiver was incorporated into the PHY implementation which is further discussed in Sect. 6. Empirical results reported therewith offer clear evidence that using DDE is quite viable in treating the channel estimate aging effect.

Finally, it is worthwhile to mention that the consistent availability of a global positioning system (GPS) signal would indeed help synchronize cooperative transmitters and thus eliminate the CCFO problem. However, it is important also to emphasize that losing the GPS signal for just a few seconds may cause transmitters' clocks to drift substantially, and therefore, the CCFO problem reemerges again. This is why it is paramount to fortify receivers with GPS-independent algorithms.

5.3 Cooperative Power Delay Profile

The power delay profile (PDP) of the cooperative channel is unique in the sense that it contains many strong yet slightly delayed signal arrivals [54]. This creates a power spectral density (PSD) shape that is also fundamentally different from that corresponding to the classical PTP channel. This is illustrated in Fig. 11. As a consequence, the PDP of the cooperative channel brings forward two PHY design challenges as explained in what follows.

5.3.1 Large Dynamic Range

Channel simulations have been carried out to characterize the dynamic range of the PSD of the cooperative channel. Results are depicted in Fig. 12 where the cumulative density function (CDF) of the PSD dynamic range is plotted for two cases, $I = 1$ and $I = 3$.

The dynamic range of the channel's spectral response dictates the dynamic range of the receiver's FFT block. This is because OFDM receivers typically employ the frequency-domain equalizers (FDE) to address the frequency selectivity of the channel. The FFT block must be able to cope with larger channel dynamic ranges. Otherwise, it will cause severe degradations in the FDE performance due to clipping, and consequently, it will adversely affect the overall receiver performance. In

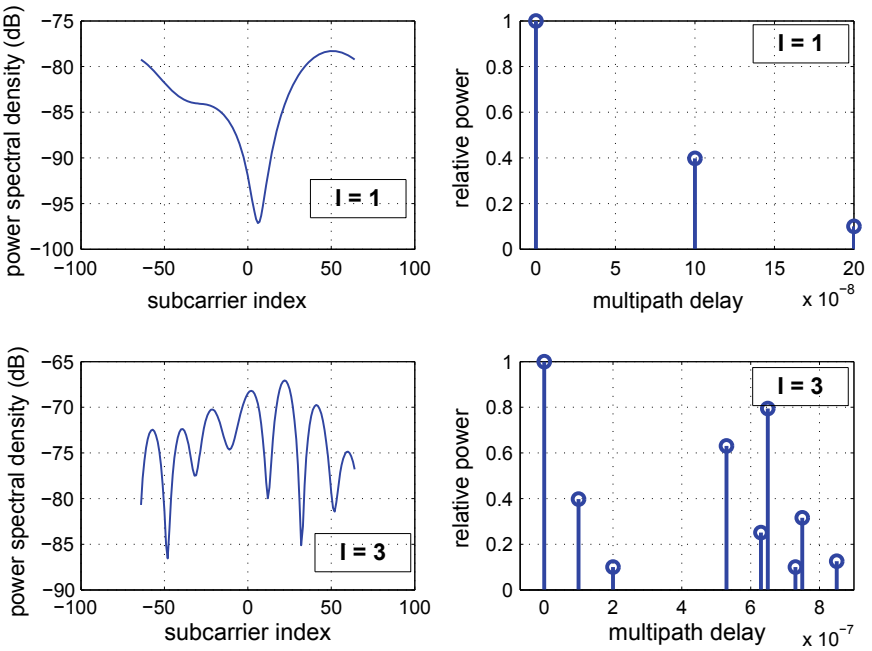


Fig. 11 The PDP and PSD of the cooperative channel (with $I = 3$) compared to that of the PTP channel. The PSD is measured over a 10 MHz channel with 128-point FFT

conclusion, the fixed-point design of the FFT block must accommodate the dynamic range requirements of cooperative transmission particularly in terms of memory resources allocated.

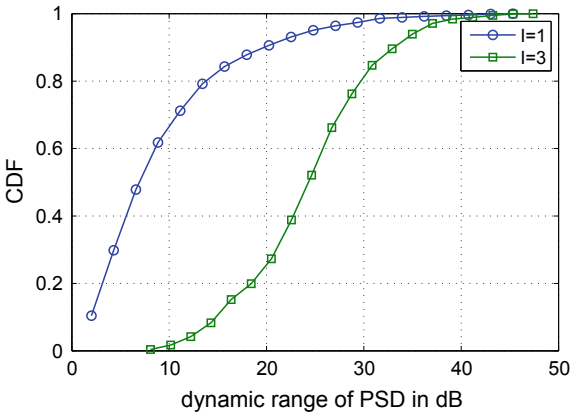


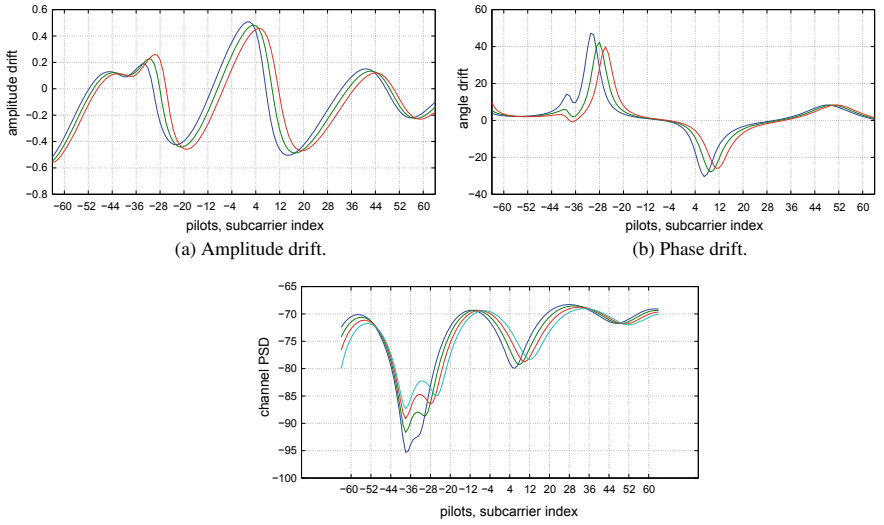
Fig. 12 The CDF of the dynamic range of the PSD. The mean dynamic range was computed to be 10.1 dB in case of one transmitter (PTP case) and 25.3 dB in case of three cooperative transmitters

5.3.2 High Frequency Selectivity

Indeed, best design practice calls for inserting pilots within the PHY frame. These pilots can be used to track phase and amplitude drifts of the channel coefficients. Pilot tones carry training symbols that are known a priori by the receiver in order to update the channel estimates. OFDM systems typically employ a comb-type pilot subcarrier arrangement whereby pilots are inserted regularly in the frequency domain [62].

Linear interpolation is mostly used to estimate channel coefficients at subcarriers between pilots. However, the frequency-domain response of the cooperative channel is quite likely not to be linear between pilots. This is further illustrated in Fig. 13. Accordingly, it is paramount to revert to nonlinear interpolation. In the implementation presented in this chapter, a three-point quadratic interpolation is carried out in accordance with [63].

Lastly, it is worthy of noting that a modern robust forward error correction (FEC) scheme is poised to address many of the challenges associated with cooperative transmission. Low-density parity-check (LDPC) codes are great candidates for this purpose [56]. The same argument applies to the use of Turbo decoders. However, it is also important to note that the remedies outlined in this section are far less demanding in terms of onboard resource utilization compared to LDPC or Turbo codes. In one instance of implementation on field programmable gate array (FPGA) platform, the inclusion of a Turbo decoder increases the resource utilization by nearly



(a) Amplitude drift.

(b) Phase drift.

(c) Power spectral density.

Fig. 13 Frequency-domain response of the cooperative channel measured at pilot tones over consecutive OFDM symbols. The subcarrier indices shown in the plots correspond to the pilots. There are eight pilots allocated within the 128-point FFT

40% compared to less than 10% for the suite of DDE, AGC update, and quadratic interpolation.

5.4 Self-localization Random Access Blocks

At the receiver, the preamble portion of the OFDM frame consists of identical replicas arriving asynchronously from I nodes. As shown in Fig. 7, the receiver aligns its time reference to the first energy arrival of the first OFDM symbol. The receiver locks to the first energy arrival of the LTS symbol, which happens to be that of the second relay in this example. The picture is fundamentally different in the RA portion where each non-empty block contains a unique signal (assuming no collisions).

RA signals are generally expected to be nonaligned in time, as illustrated in Fig. 7. Time misalignments of the RA blocks obviously occur due to the differences in propagations delays between the relays toward the receiver in concern. Hence, for some RA blocks, the FFT processing window at the receiver will not be aligned in time to the actual start of the RA signal within that block.

The effect of time offsets in OFDM systems was studied in [64]. Here, the time offset is "toward" the guard interval, i.e., the FFT window is partially applied on the guard interval. It was shown in [64] that such an offset only introduces a phase error. For this reason, OOK was chosen as a modulation scheme for convenience since it is indifferent to phase rotations. Reverting to OOK for the RA entails nearly negligible increase in the FPGA resource utilization footprint. On the other hand, the use of OOK modulation is surely associated with a 3 dB SNR penalty compared to using binary phase shift keying (BPSK), for example. Nevertheless, there is an inherent power boost on RA blocks. This is because all of the transmit RF power is focused on the RA block of choice at each transmitter. This indeed helps compensate for the SNR penalty.

On the other hand, it has already been mentioned that the reconstruction of the localization waveforms is done sequentially. Such an approach is affordable since the time budget of the localization process is not quite constrained. In other words, it is acceptable for a node to take a few seconds to adjust its position information. Therefore, dedicated FPGA resources need not be allocated for localization. Instead, available resources can be exploited opportunistically. In fact, the relaxed time constraint allows to solve the hyperbolic equations associated with the triangulation function in a more powerful microcontroller processing unit (MCU), as suggested in Fig. 6.

Finally, it is worthwhile to have a peak under the hood on how the RA can be practically implemented. For a 128-point FFT, B_Q is set at 128 tones divided equally and contiguously over two consecutive symbols. Setting and $b = 6$, and allocating 64 tones per localization block yields $B_L = 5$ blocks. For 100-symbol OFDM frames, this an overhead contribution of just 6%. If the localization capability is switched off (i.e., in case of low likelihood of GPS signal loss), the overhead goes down to less than 1%. This is tangibly better than the 14–28% incurred by ACR predecessors.

According to (20), 64 quantization tones are good enough to serve $N = 95$ nodes with an average of $I = 3$ cooperative transmitters. At the other end, each OOK-modulated localization block has 128 tones or equivalently bits. With a rate $\frac{1}{2}$ FEC, this leaves 64 bits out of which 4 can be used for parity. It can be straightforwardly shown that the remaining 60 bits are sufficient to represent the GPS position offset of a node from the sink.

On the other hand, the localization resolution is actually function of the sampling rate and the number of subcarriers in each localization block. At 40 Msps, and noting that the number of samples per block is half of that of the whole OFDM symbol, then the resolution that can be achieved is 30 m. A high-performance ADC capable of higher sampling rates is indeed slightly more expensive but—if needed—can be used to achieve better resolution.

6 Experimental Performance Evaluation Results

The main goal of the field experimentation is to validate the key building block of ACR/FACR. This is to verify that multiple transmitters induce an array gain when concurrently transmitting the same packet.

6.1 Development Platform

Off-the-shelf OFDM-based transceivers (e.g., standard-based IEEE 802.11a/g or IEEE 802.16d/e) cannot be used for experimenting with cooperative transmission schemes [56]. This is due to the fact that cooperation invokes substantial changes to the PHY and lower MAC layers. Moreover, the challenges described in Sect. 5 mandate a more robust PHY design. Hence, it was decided to build the ACR/FACR protocol stack completely from scratch so as to have a sufficient level of flexibility and control over the design process.

To that end, a compact stand-alone software-defined radio (SDR) platform was selected (Fig. 14). A complete 128-point OFDM PHY was developed entirely for this project. The PHY supports channel bandwidths from 1 to 20 MHz with ADC sampling rates up to 40 Msps. The cyclic prefix consists of 32 samples such that the total number of samples per symbol is 160. The chosen SDR is home for a 40-KLE Altera Cyclone IV FPGA, an ARM9 microcontroller architecture, and a reconfigurable radio frequency (RF) chip from Lime Microsystems. An RF amplifier from Texas Instruments was also annexed to the platform. The OFDM PHY was built on the FPGA, while the rest of the protocol stack runs on the MCU.

The original plan was to install the SDR platforms on highly mobile stations to test PHY performance. However, it was shown in Sect. 5.2 that the CCFO effect produces a channel that is much more dynamic and time-varying than that produced

Fig. 14 An SDR platform from Nuand was used to build the fully autonomous cooperative routing scheme. The platform houses a 40-KLE Altera Cyclone IV FPGA, a Cypress microcontroller unit (MCU), and a reconfigurable RF chipset from Lime Microsystems. An RF amplifier from Texas Instruments was also annexed to the platform. The OFDM PHY was built on the FPGA, while the rest of the protocol stack runs on the MCU

by Doppler spread, even at high speeds. A corollary to this is that empirical results collected from the field under CCFO with *stationary* nodes are sufficient to ensure the implementation will successfully handle mobility. The key parameters concerning the underlying PHY design are reported in Table 3.

Table 3 Key OFDM PHY design parameters

Channel bandwidth	1–20 MHz
Frequency spectrum	0.3–3.8 GHz
Maximum RF transmit power	10 dBm
Antenna gain	3 dBi
FFT size	128 points
Maximum sampling rate	40 Msps
Preamble length (STS+LTS)	768 samples
Number of pilots	8
Turnaround time	180 μs
Useful symbol length	128 samples
Cyclic prefix length	32 samples

6.2 Equalizer Performance

The performance of the DDE implementation was investigated under a controlled setup. A dedicated BladeRF board was configured to feed three other boards with two common signals: clock and trigger, as shown in Fig. 15. The latter is used to instruct the three relays to commence the transmission of a frame that is prestored on the FPGA. The CFO is invoked locally at each transmitting node via a command line interface (CLI) utility. Similarly, each node may be configured to introduce a fixed delay after the rising edge of the trigger signal. This can be used to produce the

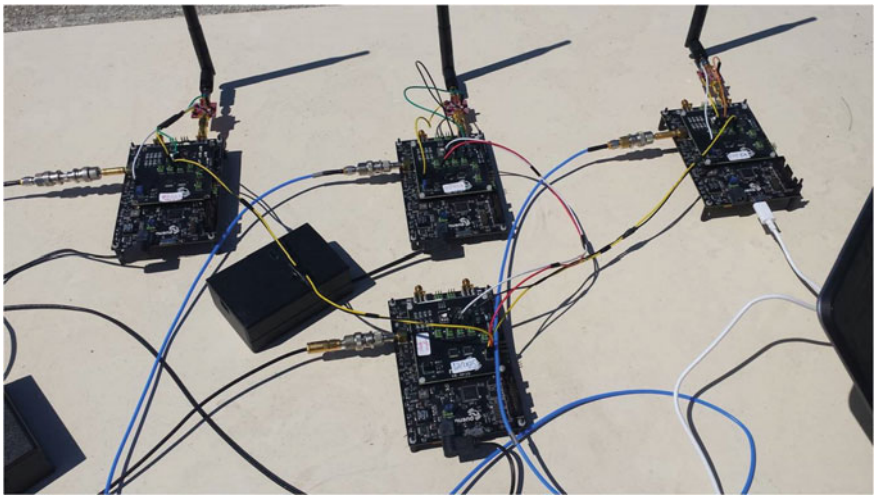


Fig. 15 Common clock and trigger signals are fed into the boards. The controlled test setup is used to measure the performance of the decision-directed equalization method

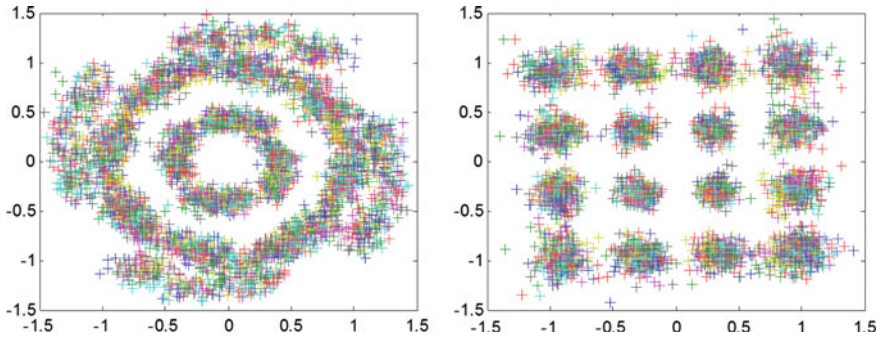


Fig. 16 DDE performance investigated for the case of concurrent transmission from three nodes with CFO1 = 1000 Hz, CFO2 = 0 Hz, CFO3 = −100Hz. Here, the received I/Q symbols are plotted

Table 4 DDE performance results

	Traditional equalizer	Decision-driven equalizer (DDE)
FER	81%	5%
EVM	−6 dB	14 dB
Highest modulation	BPSK	16 QAM

desired delay spread for the composite channel. In other words, it helps control the propagation delays $T_1^{'}, \ldots, T_I^{'}$ shown in Fig. 8.

In this test, the three transmitters were placed 12 m from the receiver. The CFO values for the transmitters were set at −100, 500, and 1000 Hz. A 16 quadrature amplitude modulation (16-QAM) modulation scheme with FEC rate of 1/2 was used. As expected, DDE is quite a viable tool to equalize highly time-varying channels. Run over a large number of iterations, the average frame error rate (FER) plunged from 81% down to less than 5% when the DDE module was activated at the receiver. The error vector magnitude (EVM) of the baseband inphase/quadrature (I/Q) stream was measured on randomly selected subset of frames in MATLAB (Fig. 16). The average EVM ascended from as low as −6 to 14 dB. Neglecting transmitter I/Q imperfections, the EVM is known to be tightly related to the receiver SNR. As such, the DDE module can be said to offer a gain of nearly 20 dB while introducing less than 5% in the overall footprint of the PHY code. These results are summarized in Table 4.

Table 5 Results from the array gain test

	SNR (dB)	RSSI (dBm)	FER (%)	reach (m)
Tx_1 only	16.1	−88.2	2.9	70
	11.2	−93.1	50.8	115
Tx_2 only	16.5	−87.5	0.8	65
	11.8	−94.1	76.2	115
Tx_3 only	16.4	−88.0	1.1	65
	12.4	−93.5	47.0	115
All three	16.3	−86.7	2.0	115

6.3 Array Gain

The goal of this test case was to measure the array gain as well as the maximum reach gain that can be obtained by means of autonomous cooperative transmission. The same setup presented in the previous subsection was used. To obtain the maximum reach gain, the CFO was forced to zero on all three transmitters. The three transmitters were always kept equidistant from the receiver. The test was carried out in an open space parking lot surrounded by light vegetation. All nodes were placed one meter above ground level. Results of this test are reported in Table 5. Each result corresponds to an average value taken over an ensemble of 10,000 frames.

The first stage of this stage was to measure the communication range for each individual transmitter. The communication range here was defined as the maximum reach such that an average FER target of $\leq 3\%$ is maintained. The receiver was gradually moved away in steps of 5 m. As reported in Table 5, the communication range was around 65–70 m. The slight discrepancy in results is due to the different multipath channels since transmitters are not co-located. Another factor is the approximate nature of any method for computing the SNR on the preamble signal.

Next, the communication range for the case of three cooperative transmitters was measured by gradually moving the receiver away in steps of 5 m. The maximum range was measured to be 115 m, i.e., the reach gain was 50 m or equivalently 77%. Indeed, the reach gain highly depends on the propagation characteristics, which in return relates to the environment where the test is performed.

Now in order to characterize the array gain, each transmitter was placed 115 m away from the receiver and the SNR was measured. The array gain is computed here as the difference between the SNR obtained under cooperative transmission and the average of individual SNR values. It is clear from Table 5 that autonomous cooperative transmission is able to offer nearly 4.5 dB of array gain. This result is quite interesting since it is very close to the theoretical maximum array gain with three transmitters, i.e., $10 \log 3 = 4.77$ dB.

7 Conclusions

There is a growing trend for streaming live vision-based data from the field to enhance visibility and assist decision-making during mission-critical situations. The dissemination of live vision-based data feeds demands high bandwidth in addition to low latency.

Over-the-counter wireless technologies available today have been shown to fall short in meeting the demands of next-generation mission-critical applications. As such, mobile ad hoc networking (MANET) has resurfaced again as a viable contender in place of Wi-Fi and LTE. Having said that, classical path-oriented MANET routing techniques are notoriously known to accumulate substantial protocol overhead as the network grows in scale. Subsequently, it has been shown that autonomous cooperative routing (ACR) is well positioned to meet the goals and requirements of mission-critical operations.

To that end, the implementation of ACR on commercial hardware platforms entails a few practical challenges which have not been quite addressed in literature. The foremost challenge concerns the receiver's capability in handling the aggressive nature of the cooperative wireless channel. The cooperative channel is highly time-varying therefore causing channel estimates to become obsolete pretty quickly. A robust channel equalizer based on the use of decision-drive estimation (DDE) was presented to remedy this issue. On the other hand, the cooperative channel has been shown to feature a high level of selectivity in the spectral domain which was handled by means of optimized pilot signal processing. The chapter also presented a fully astomous version of ACR. Practical implementation considerations have been also highlighted offering some evidence of the advents of full autonomy.

Finally, the chapter presented an experimental setup that was developed specifically to validate the basic building blocks of ACR/FACR. A protocol stack was built from scratch for that purpose. Field experiments were carried out and were able to validate some of the performance enhancement propositions outlined in the various sections of the chapter.

References

1. Suriyachai, P., Roedig, U., Scott, A.: A survey of mac protocols for mission-critical applications in wireless sensor networks. IEEE Commun. Surv. Tutor. **14**(2), 240–264 (2012)
2. Fink, J., Ribeiro, A., Kumar, V.: Robust control of mobility and communications in autonomous robot teams. IEEE Access **1**, 290–309 (2013)
3. Ghafoor, S., Sutton, P.D., Sreenan, C.J., Brown, K.N.: Cognitive radio for disaster response networks: survey, potential, and challenges. IEEE Wirel. Commun. **21**(5), 70–80 (2014)
4. Akyildiz, I., Melodia, T., Chowdhury, K.: A survey on wireless multimedia sensor networks. Elsevier J. Comput. Netw. **51**(4), 921–960 (2007)
5. Zhang, Z.J., Lai, C.F., Chao, H.C.: A green data transmission mechanism for wireless multimedia sensor networks using information fusion. IEEE Wirel. Commun. **21**(4), 14–19 (2014)
6. Yang, L., Yang, S.-H., Plotnick, L.: How the Internet of Things technology enhances emergency response operations. Technol. Forecast. Soc. Change Elsevier **80**(9), 1854–1867 (2013)

7. Chai, P.R.: Wearable devices and biosensing: Future frontiers. J. Med. Toxicol. 1–3 (2016)
8. Panayides, A., Antoniou, Z.C., Mylonas, Y., Pattichis, M.S., Pitsillides, A., Pattichis, C.S.: High-resolution, low-delay, and error-resilient medical ultrasound video communication using h.264/avc over mobile wimax networks. IEEE J. Biomed. Health. Inform. **17**(3), 619–628 (2013)
9. Bergstrand, F., Landgren, J.: Using live video for information sharing in emergency response work. Int. J. Emerg. Manag. **6**(3–4), 295–301 (2009)
10. Blair, A., Brown, T., Chugg, K.M., Halford, T.R., Johnson, M.: Barrage relay networks for cooperative transport in tactical manets. In MILCOM 2008—2008 IEEE Military Communications Conference, Nov 2008, pp. 1–7
11. Bergstrand, F., Landgren, J.: Visual reporting in time-critical work: exploring video use in emergency response. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services, pp. 415–424 (2011)
12. Nunes, D.S., Zhang, P., Sa Silva, J.: A survey on human-in-the-loop applications towards an internet of all. IEEE Commun. Surv. Tutor. **17**(2), 944–965 (2015)
13. Felts, R., Leh, M., McElvaney, T.: Public safety analytics r&d roadmap. National Institute of Standards and Technology (NIST), U.S. Department of Commerce, Technical Note 1917, Apr 2016
14. Bayezit, I., Fidan, B.: Distributed cohesive motion control of flight vehicle formations. IEEE Trans. Ind. Electron. **60**(12), 5763–5772 (2013)
15. Berni, J., Zarco-Tejada, P.J., Suarez, L., Fereres, E.: Thermal and narrowband multispectral remote sensing for vegetation monitoring from an unmanned aerial vehicle. IEEE Trans. Geosci. Remote Sens. **47**(3), 722–738 (2009)
16. Siebert, S., Teizer, J.: Mobile 3d mapping for surveying earthwork projects using an unmanned aerial vehicle (uav) system. Autom. Constr. Elsevier **41**, 1–14 (2014)
17. Lin, X., Andrews, J.G., Ghosh, A., Ratasuk, R.: An overview of 3g pp device-to-device proximity services. IEEE Commun. Mag. **52**(4), 40–48 (2014)
18. Bader, A., Ghazzai, H., Kadri, A., Alouini, M.S.: Front-end intelligence for large-scale application-oriented internet-of-things. IEEE Access **4**, 3257–3272 (2016)
19. Carl, L., Fantacci, R., Gei, F., Marabissi, D., Micciullo, L.: Lte enhancements for public safety and security communications to support group multimedia communications. IEEE Netw. **30**(1), 80–85 (2016)
20. Gharbieh, M., ElSawy, H., Bader, A., Alouini, M.-S.: Tractable stochastic geometry model for iot access in lte networks. In: To Appear in Proceedings of IEEE Globecom 2016, Washington D.C., December 2016
21. Kiess, W., Mauve, M.: A survey on real-world implementations of mobile ad-hoc networks. Ad Hoc Netw. **5**(3), 324–339 (2007)
22. Bellalta, B.: IEEE 802.11ax: high-efficiency wlans. IEEE Wirel. Commun. **23**(1), 38–46 (2016)
23. Abouzeid, A.A., Bisnik, N.: Geographic protocol information and capacity deficit in mobile wireless ad hoc networks. IEEE Trans. Inf. Theory **57**(8), 5133–5150 (2011)
24. Request for information, novel methods for information sharing in large scale mobile ad-hoc networks, defense advanced research projects agency (darpa), darpa-sn-13-35, April 2013
25. Halford, T.R., Chugg, K.M., Polydoros, A.: Barrage relay networks: system and protocol design. In: 21st Annual IEEE International Symposium on Personal, pp. 1133–1138. Sept, Indoor and Mobile Radio Communications (2010)
26. Halford, T.R., Chugg, K.M.: Barrage relay networks. In: Information Theory and Applications Workshop (ITA), 2010, Jan 2010, pp. 1–8
27. Acer, U.G., Kalyanaraman, S., Abouzeid, A.A.: Weak state routing for large-scale dynamic networks. IEEE/ACM Trans. Netw. **18**(5), 1450–1463 (2010)
28. Halford, T.R., Chugg, K.M.: The stability of multihop transport with autonomous cooperation. In: 2011—MILCOM 2011 Military Communications Conference, Nov 2011, pp. 1023–1028
29. Bader, A., Abed-Meraim, K., Alouini, M.-S.: An efficient multi-carrier position-based packet forwarding protocol for wireless sensor networks. IEEE Trans. Wirel. Commun. **11**(1), 305–315 (2012)

30. Lakshmi, V., Thanayankizil, A.K., Ingram, M.A.: Opportunistic large array concentric routing algorithm (olacra) for upstream routing in wireless sensor networks. Ad Hoc Netw. **9**(7), 1140–1153 (2011)
31. Ke, C.-K., Chen, Y.-L., Chang, Y.-C., Zeng, Y.-L.: Opportunistic large array concentric routing algorithms with relay nodes for wireless sensor networks. Comput. Electr, Eng (2016)
32. Halford, T.R., Courtade, T.A., Turck, K.A.: The user capacity of barrage relay networks. In: MILCOM 2012—2012 IEEE Military Communications Conference, Oct 2012, pp. 1–6
33. Xiang, X., Wang, X., Zhou, Z.: Self-adaptive on-demand geographic routing for mobile ad hoc networks. IEEE Trans. Mob. Comput. **11**(9), 1572–1586 (2012)
34. Intelligent transport systems (its); vehicular communications; geonetworking; part 4: geographical addressing and forwarding for point-to-point and point-to-multipoint communications; sub-part 1: Media-independent functionality, v1.2.0, Oct 2013
35. Sanchez, J., Ruiz, P., Marin-Perez, R.: Beacon-less geographic routing made practical: challenges, design guidelines, and protocols. IEEE Commmun. Mag. **47**(8), 85–91 (2009)
36. Scaglione, A., Goeckel, D., Laneman, J.: Cooperative communications in mobile ad hoc networks. IEEE Signal Process. Mag. **23**(5), 18–29 (2006)
37. Sirkeci-Mergen, B., Scaglione, A.: Randomized space-time coding for distributed cooperative communication. ICC (2006)
38. Sirkeci-Mergen, B., Scaglione, A.: Randomized space-time coding for distributed cooperative communication. IEEE Trans. Signal Process. **55**(10), 5003–5017 (2007)
39. Sharp, M., Scaglione, A., Sirkeci-Mergen, B.: Randomized cooperation in asynchronous dispersive links. IEEE Trans. Commun. **57**(1), 64–68 (2009)
40. Li, Y., Zhang, Z., Wang, C., Zhao, W., Chen, H.-H.: Blind cooperative communications for multihop ad hoc wireless networks. IEEE Trans. Veh. Technol. **62**(7), 3110–3122 (2013)
41. Brian, R.H., Hwang, G.: Barrage relay networks for unmanned ground systems. In: Military Communications Conference, 2010—MILCOM 2010, Oct 2010, pp. 1274–1280
42. Lee, D.K., Chugg, K.M.: A pragmatic approach to cooperative communication. In: MILCOM 2006—2006 IEEE Military Communications Conference, Oct 2006, pp. 1–7
43. Bader, A., Alouini, M.-S.: An ultra-low-latency geo-routing scheme for team-based unmanned vehicular applications. In: IEEE Globecom Workshops (GC Wkshps), 2015. IEEE, pp. 1–6 (2015)
44. Bader, A., Alouini, M.S.: Localized power control for multihop large-scale internet of things. IEEE Internet of Things J. **3**(4), 503–510 (2016)
45. Zorzi, M., Rao, R.: Geographic random forwarding (GeRaF) for ad hoc and sensor networks: multihop performance. IEEE Trans. Mob. Comput. **2**(4), 337–348 (2003)
46. Gupta, P., Kumar, P.R.: The capacity of wireless networks. IEEE Trans. Inf. Theory **46**(2), 388–404 (2000)
47. Bisnik, N., Abouzeid, A.A.: Queuing network models for delay analysis of multihop wireless ad hoc networks. Ad Hoc Netw. Elsevier **7**(1), 79–97 (2009)
48. Bisnik, N., Abouzeid, A.A.: Queuing delay and achievable throughput in random access wireless ad hoc networks. In: 2006 3rd Annual IEEE Communications Society on Sensor and Ad Hoc Communications and Networks, Sept 3, pp. 874–880 (2006)
49. Wu, H., Peng, Y., Long, K., Cheng, S., Ma, J.: Performance of reliable transport protocol over ieee 802.11 wireless lan: analysis and enhancement. In: INFOCOM 2002. Twenty-First Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings, vol. 2, pp. 599–607. IEEE (2002)
50. Bader, A., Abed-Meraim, K., Alouini, M.S.: Reduction of buffering requirements: Another advantage of cooperative transmission. IEEE Sens. J. **15**(4), 2017–2018 (2015)
51. McDonald, A.B., Znati, T.F.: A mobility-based framework for adaptive clustering in wireless ad hoc networks. IEEE J. Sel. Areas Commun. **17**(8), 1466–1487 (1999)
52. Squartini, T., Picciolo, F., Ruzzenenti, F., Garlaschelli, D.: Reciprocity of weighted networks. Scientific reports, vol. 3, 2013
53. Eriksson, M., Mahmud, A.: Dynamic single frequency networks in wireless multihop networks—energy aware routing algorithms with performance analysis. In: Proceedings of

The 10th IEEE International Conference on Computer and Information Technology, Bradford, UK, pp. 400–406, May 2010
54. Bader, A., Alouini, M.S.: Mobile ad hoc networks in bandwidth-demanding mission-critical applications: practical implementation insights. IEEE Access **5**, 891–910 (2017)
55. Srinivasa, S., Haenggi, M.: Distance distributions in finite uniformly random networks: theory and applications. IEEE Trans. Veh. Technol. **59**(2), 940–949 (2010)
56. Qiu, H., Wang, K., Psounis, K., Caire, G., Chugg, K.M.: High-rate wifi broadcasting in crowded scenarios via lightweight coordination of multiple access points. In: MobiHoc '16 Proceedings of the 17th ACM International Symposium on Mobile Ad Hoc Networking and Computing, July 2016, pp. 301–310
57. Rappaport, T.: Wireless Communications: Priciples and Practice, 2nd edn. Prentice Hall (2001)
58. Schulze, H., Lueders, C.: Theory and Applications of OFDM and CDMA, 1st edn. John Wiley and Sons Ltd (2005)
59. Zhao, B., Valenti, M.C.: Practical relay networks: a generalization of hybrid-ARQ. IEEE J. Sel. Areas Commun. **23**(1) (2005)
60. Rouphael, T.J.: Wireless Receiver Architectures and Design: Antennas, RF. Mixed Signal, and Digital Signal Processing. Elsevier, Synthesizers (2014)
61. Ran, J., Grunheid, R., Rohling, H., Bolinth, E., Kern, R.: Decision-directed channel estimation method for OFDM systems with high velocities. In: Vehicular Technology Conference, 2003. VTC 2003-Spring. The 57th IEEE Semiannual, April 2003, vol. 4, pp. 2358–2361
62. Hsieh, M.-H., Wei, C.-H.: Channel estimation for OFDM systems based on comb-type pilot arrangement in frequency selective fading channels. IEEE Trans. Consum. Electron. **44**(1), 217–225 (1998)
63. Coleri, S., Ergen, M., Puri, A., Bahai, A.: Channel estimation techniques based on pilot arrangement in ofdm systems. IEEE Trans. Broadcast. **48**(3), 223–229 (2002)
64. Athaudage, C.: BER sensitivity of OFDM systems to time synchronization error. In: The 8th International Conference on Communication Systems (ICCS'02), Singapore, vol. 1, pp. 42–46, Nov 2002

Using Models for Communication in Cyber-Physical Systems

Yaser P. Fallah

Abstract One of the main components of cyber-physical systems (CPS) is the underlying communication mechanism that enables control and decision-making. Communication has traditionally taken the form of sensing a physical phenomenon, or a cyber process, and then transmitting the sensed data to other entities within the system. With CPS being in general much more complex than a single physical or cyber process, the requirements on communication and data content are high. Therefore, communication of all the required information for control of a CPS may become a challenge. In this chapter, we present a new paradigm in communication which utilizes communication of models and model updates rather than raw sensed data. This approach, which transforms overall communication structure, has the potential to considerably reduce the communication load, and provide a mechanism for richer understanding of the processes whose data is being received over a communication link. We take the example of a vehicular CPS that relies on communication for collision avoidance and demonstrate the effectiveness of the model-based communication (MBC) concept.

1 Introduction

Communication is one of the main aspects of cyber-physical networked systems and sensor networks. The role of communication in such systems is to transfer knowledge of a particular phenomenon from one system component to another. Such knowledge is usually presented in the form of signals or samples of physical or cyber phenomena in a CPS. Therefore, communication systems are structured in the form of sampling physical or cyber phenomena and then transmitting the sampled data over a link. The receivers of data are expected to know the meaning of the communicated samples, and after estimation and correction, the received data is interpreted using preexisting knowledge of the phenomena to which they belong. For example, when temperature

Y. P. Fallah (✉)
University of Central Florida, Orlando, USA
e-mail: yaser.fallah@ucf.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_3

value is sensed at a specific time and location, and transmitted in a sensor network, the receiver is expected to know how to interpret temperature data and estimate it for missing samples. While sensed data is simple in this and many other cases, in some other scenarios, it may represent much more complicated dynamics. An example is the movement of a vehicle. Representing movement data usually happens through high rate sampling of several parameters, such as position, velocity, heading, acceleration, etc., and communicating them to receivers which are expected to use certain models to reconstruct the vehicle movement. The existing knowledge of the models that describe system dynamics (in this case in the form of kinematic equations) and frequent sampling and communication of the dynamics are two of the main characteristics of networked systems with relatively high-speed dynamics. For example, many vehicle safety or control applications require position data with less than 1 m of error (95 percentile). The requirements are even higher when automated applications are considered. The precision requirements on knowing the movement of a subject vehicle (or a tracked dynamic system) determine the requirements on communication and on model fidelity.

In traditional communication structure, models are already used at the receiver side for the purpose of estimation and reconstruction of the original signal. In such systems, the assumption is that the receiver has knowledge of a model that represents the source of communicated data and uses this knowledge to correct issues that may exist in communication. For example, when position data of a vehicle is communicated, missing position samples may be estimated using models of vehicle movement. On the other hand, these same models may be used at the sender to selectively sample and communicate data points that are more valuable for estimation purposes [1, 2]. In all these designs, the assumption is that the sender and receiver are aware of the models that represent the system of interest.

Overall, the communication system design in CPS, which is almost always in discrete form, has to consider three aspects: how/when to sample a signal and transmit it, how to estimate the source signal using the received samples (which may be a subset of transmitted samples), and how to transfer data over a link. In this chapter, we only discuss the first two aspects **of sampling/communication** and **estimation**; the issue of how data is transferred over a link is a pure communication question and is not specific to CPS. The sampling and estimation aspects are often dependent on the specific CPS application or system. The challenge of designing these two components is the main focus of discussion in this chapter.

To better describe the role of models in communication, we elaborate on the example of vehicle tracking and applications such as cooperative collision warning (CCW) here [3, 4]. CCW systems rely on vehicles broadcasting their position information over a local wireless network, as in Fig. 1. Currently, dedicated short range communication or DSRC [5] technology is being considered for establishing the wireless connections. A vehicle that receives information from other vehicles in its vicinity creates a real-time position map of all vehicles in its surrounding. The map is built by estimating the position from received data. Communication and estimation modules are the building blocks of the situational awareness component (Fig. 2). The map is analyzed continuously to identify hazardous situations. Obviously, when

Fig. 1 Cooperative safety systems, such as CCW, rely on the broadcast of vehicle movement information over a local wireless network

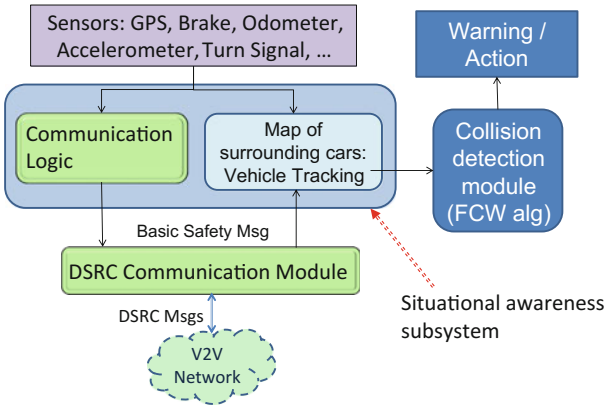


Fig. 2 A block diagram depiction of cooperative safety systems; communication and vehicle tracking modules constitute the situational awareness subsystem

communication loss happens, estimation of other vehicles positions and identifying hazards become unreliable.

2 Using Communication for Tracking a Vehicle (or a System)

Vehicle tracking in CCW relies on each vehicle sampling its movement data such as position, speed, and acceleration on a regular basis (usually at 10 Hz) and communicating it over a local wireless network. The communicated data is a snapshot of the sender vehicle's *state* at the sampling time. While DSRC-based communicated messages usually contain some additional information such as path history (refer to

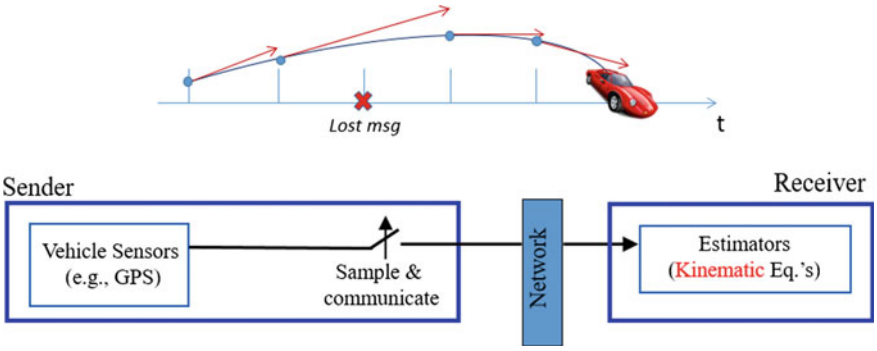


Fig. 3 Kinematic models can be used for estimation of the sender state when samples are not available or between sample times

standards such as SAE J2735 for details [6]), in this chapter, we focus on current state data to better explain the communication concepts. A receiver uses the received snapshots of the state information to reconstruct and estimate the sender's state at all times. When some samples are lost, or for the times in between received samples, models of the sender will be used to estimate its state at any given time (see Fig. 3).

Examples of vehicle models include first-order kinematic model (constant speed model) and second-order kinematic model (constant acceleration model), which are simple equations relating acceleration, speed, and position. Other estimation and tracking approaches are also possible, for example, using Kalman filter. The common factor in estimation is the presence of some form of a model that is an accurate enough representation of the vehicle dynamics (or the dynamic system that is being tracked). Accuracy requirements for the model will be dictated by the application and may vary widely.

The use of estimators requires the receiver to have access to the model, whereas the sender may actually not need to know the model and only sample the state of the system. However, some implicit knowledge of the model is assumed at the sender, since relevant state variables (that drive the model at the receiver) need to be sampled and communicated. The basic method of communication, which is used in some early prototypes of cooperative safety systems, operates at a rate of 10 Hz (or 5 Hz) and includes broadcast of movement data (such as position, heading, speed, and acceleration) in DSRC basic safety messages (BSM). We call this the "*baseline*" method. Receivers usually utilize a constant speed or constant acceleration model to reconstruct sender's trajectory.

Senders' awareness of the model and the receiver estimation process can be taken one step further. A sender can use the same estimation models that are utilized by receivers and locally reproduce a copy of the estimated state of the sender at the receiver. This reproduced estimation can be compared to the actual signal at the sender, in order to determine what the estimation error may be at the receiver at any

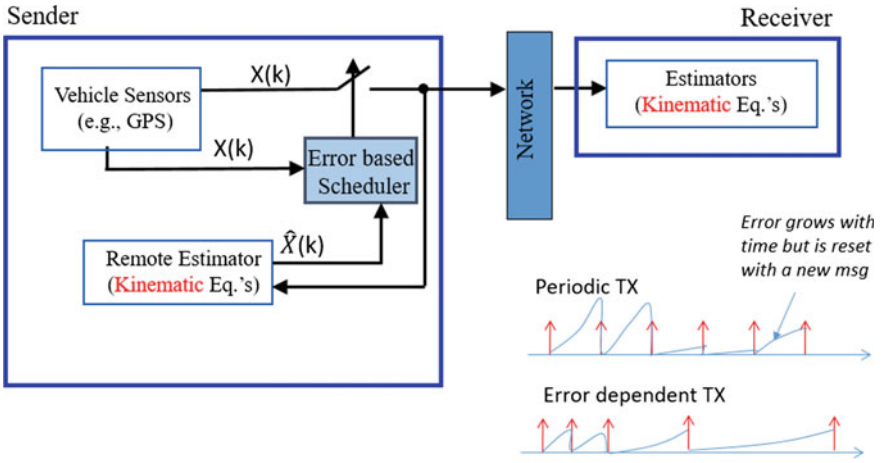


Fig. 4 Error-dependent communication logic: recreating receiver's estimator model in the sender allows for determining when a new sample needs to be transmitted

given point in time. Such error may then be used to decide when new samples need to be taken or communicated to the receiver. This method, which is called *error-dependent* (ED) communication logic and depicted in Fig. 4, is in fact the basis of the mechanism presented in [1] and adopted by SAE J2945/1 standard [7].

One obvious shortcoming of the structure in Fig. 4 is that the loss of information in the network will impact the result of estimation in the receiver, but not in the sender. If the loss of communicated data can be determined, i.e., the sender knows which message was lost, the same loss can be applied to the input of the estimator in the sender. However, determining the loss of packets is not always possible. In particular, in broadcast wireless networks with no acknowledgement (e.g., the DSRC networks for vehicular safety applications), there is no easy mechanism for the sender to identify lost messages at each receiver. To overcome this issue, the work in [1] proposes to use a metric such as packet error rate (PER) and randomly drop the packets that are input to the estimator in the sender. The overall effect of this simulated packet loss will be similar to what the receivers experience. This structure is shown in Fig. 5.

Using error-dependent communication logic, it is possible to reduce the communication load of applications such as CCW by a factor of 2–3 [8]. The main enabling concept of the ED logic is the use of models of the dynamic system (vehicle) in both receiver and the sender. The idea of *model-based communication* (MBC, [9]) transforms this concept to a new level, in which models are themselves also part of the communicated content. With communication of models, there is no need for the receiver to know the dynamics of the system a priori, since the sender will provide the model and its updates in real time.

An important advantage of this feature is that models used for estimation do not need to be very general and fit all possible behaviors of the dynamic system at

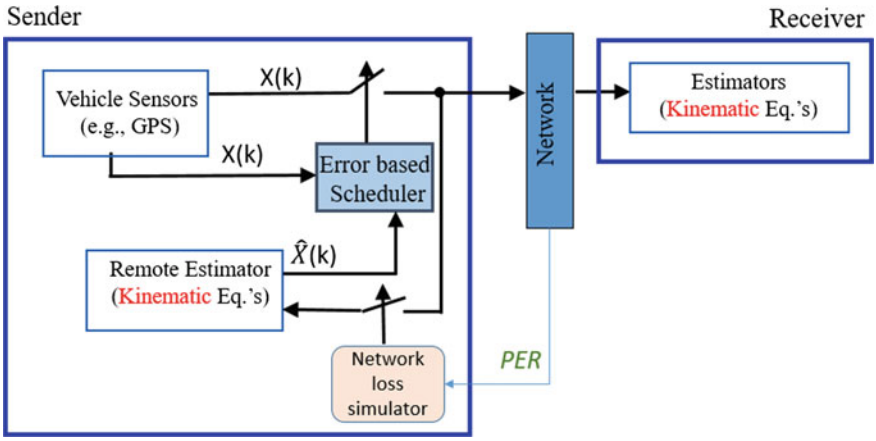


Fig. 5 Adding a network loss simulator to the sender allows for a better approximation of estimation error at the receiver

all times. The communicated models themselves could be updated and learnt at the sender and adjusted as the system evolves. The sender can always provide the receiver with updates or changes to the model. This feature also allows for considerable reduction in the communication load, as more complex models or more precise models can now be used without adding much to the overhead of communication.

To demonstrate how this concept works, the CCW structure can be imagined to use more complex models of vehicle movement instead of raw movement parameters. Figure 6 depicts this vision. Here, each vehicle occasionally communicates its movement model (e.g., at intervals in the order of tens of seconds), and more frequently broadcasts updates to the model on state transitions, parameter changes or model input changes. Using the ED logic is also possible here, and it is expected that ED logic further reduces the communication load [9].

To fully realize MBC and its benefits, viable models of the dynamic systems of interest need to be developed. For the case of vehicle movement, the simple kinematic equations need to be extended to models that can predict movement parameters for longer horizons than few hundred milliseconds. Modeling and predicting vehicle movement is a subject of its own and not the focus of this book. There are many ongoing research activities that focus on this subject; readers interested in this topic may refer to [10] and the references therein. Here, for the purpose of explaining the MBC concept, we use some established car following models that are used in traffic simulators to model driver and vehicle behavior (on an average basis). Later in this chapter, a short discussion on other methods of modeling vehicle movement and their implications for MBC are discussed.

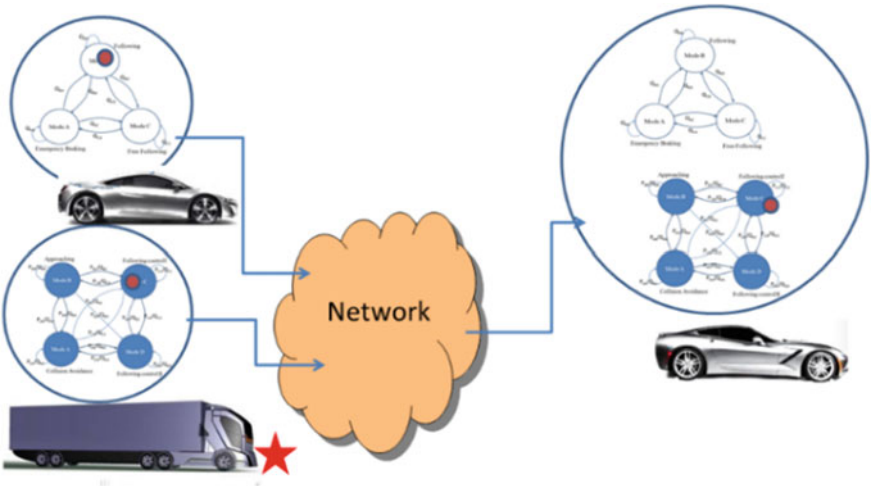


Fig. 6 Model-based communication relies on exchange of models and model updates instead of raw sample data. Hypothetical multiple-mode models are used by each sender

3 Example of Models for Vehicle Movement

Car following models [11] have been developed for the purpose of simulating vehicle movement in microscopic traffic simulators. These models capture movement dynamics of a vehicle. These dynamics depend on many factors such as driver behavior, dynamics of the vehicle itself, geometry of the road, traffic conditions, etc. It is in general easier to model automated vehicles since their movements follow known procedures, and human drivers add a significant level of uncertainty that is not easy to model. The car following models that have been developed for simulation abstract all uncertainties resulting from environment conditions and driver behavior and provide mathematical forms that express vehicle movement in a few predefined modes [12–14]. Given the wide variety of situations and driver behaviors that are possible, the car following models tend to be too abstract. One method to achieve higher fidelity in modeling is by expanding the possible modes, or tuning (training) models to drivers and situations. Machine learning techniques are very useful in such cases [10]. Communicating the time-changing models is intrinsically handled in the MBC approach.

To see how MBC can adopt models, we start with the car following model from [11]. The movement model from [11] is a relatively simple model that has been used in simulation tools such as MITSIM (other tool such as VISSIM or SUMO may use different models). This model describes how a human-operated car reacts to changes in speed or velocity of the car in front. The reaction is modeled in the form of the acceleration of a following vehicle (FV), in response to the known speed and acceleration of a vehicle that is in front of it (called lead vehicle, LV). It is assumed

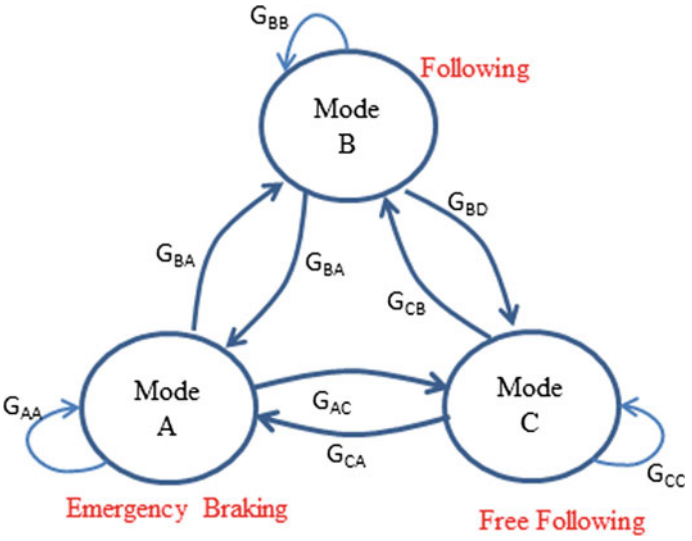


Fig. 7 Example of a 3-mode model of vehicle movement presented in a variation of hybrid automaton format

that the speed and acceleration values of LV are available through either vehicle sensors, DSRC communication, or driver's visual sensing.

The model in [11] defines three *modes* of driving behavior, with equations describing the vehicle acceleration in each mode. One way to describe this model is to use a Hybrid System (HS) form, for example, as a Hybrid Automaton [15] (or in more general term as a Hybrid I/O Automaton [16]). If transitions or other state information take probabilistic forms then probabilistic (or stochastic) variations of hybrid representation [17, 18] need to be considered. Figure 7 depicts a 3-mode HS in a variation of hybrid automata form. Each mode represents a specific driving behavior; the equations for the acceleration a in each mode of the HS are as follows:

Mode A (Emergency braking):

$$a = \begin{cases} \min(a^+, a_{LV} - 0.5(v - v_{LV})^2/g\} & v > v_{LV} \\ \min(a^+, a_{LV} + 0.25a^-\} & v \leq v_{LV} \end{cases} \tag{1}$$

Mode B (following):

$$a = \sigma^{\pm}.\frac{v^{\beta^{\pm}}}{g^{\gamma^{\pm}}}(v_{LV} - v) \tag{2}$$

β, γ, and σ are model parameters, $+$ or $-$ indicate accelerating ($v < v_{LV}$) or decelerating ($v > v_{LV}$) situations, where v is the current speed of FV, and LV refers to the leading vehicle.

Mode C (free flowing):

$$a = \begin{cases} a^{+} if\ v < V_{target} \\ 0 if\ v = V_{target} \\ a^{-} if\ v > V_{target} \end{cases} \quad (3)$$

a^{+} or a^{-} are typical max and min accelerations for a car. These values are reported in [11]. V_{target} is the target speed of a car (e.g., road speed limit) if there is no other vehicle immediately in front of it. In the last mode, it is expected that the driver smoothly adjusts vehicle speed until reaching the target speed, which may be the speed limit.

The model is usually evaluated at a fixed frequency (for example every 100 ms) and the mode appropriate for the situation is selected. Within each mode, the vehicle will move with the calculated acceleration value using a constant acceleration model. Transition between modes can happen due to the guard conditions being satisfied or with probabilities in probabilistic models. The transition conditions (or probabilities) are denoted in Fig. 7 as G_{ij} (for transition from mode i to j). The transition conditions are evaluated based on environmental inputs, for example, from the parameters of the vehicle in front (LV) or the situation of the FV. This is a natural choice since a driver usually remains in the same mode until the traffic situation or the situation of the following vehicle (FV) changes.

4 Using Models in MBC

The model in Fig. 7 can be used in MBC in several different ways. One of the main principles in MBC allows for the entire model and its parameters to be communicated. This is usually a step that happens occasionally and with low frequency. For example, consider the vehicles in Fig. 8. Here our focus is on the last vehicle denoted as host vehicle (HV). We do not explicitly use the LV and FV notation in this example, since this notation can be used for any pair of vehicles, and there are several LVs and FVs in this figure. The remote vehicle (RV) is the vehicle that is being tracked by HV. For the sake of our discussion on MBC, assume that RV us moving according to the car following model of Fig. 7. This model will be communicated to HV a short while after HV comes into communication range of RV. Before receiving this 3-mode model, a default kinematic model is used for tracking RV at HV. It is therefore necessary that model updates and some basic information are included in the frequent messages of RV that allow it to be tracked at least using the default kinematic models.

Following reception of the 3-mode model of RV in HV, tracking in HV will become considerably more precise since acceleration values of RV can now be predicted for a short time horizon in the future. For this purpose, inputs to the movement model of RV (for example the speed of the vehicle in front of RV) can be used as model updates. Such information may already be available at HV, since in a DSRC network, HV

Fig. 8 Example of host and remote vehicles: HV and RV

will hear the messages from the vehicle that is in front of RV too. Nevertheless, that information can be assumed to be included in "model updates" that are sent by RV. Using the model and model updates, the tracking process in HV will in fact becomes an emulation of RV movement using the available model information. Methods such as the error-dependent logic described in previous section can be utilized here to determine when and what model updates are needed.

An error-dependent model for MBC will operate at different levels. At the highest level, the entire model (such as the HS in Fig. 7) will be transmitted occasionally and only when models completely change. For example, when a vehicle exists a highway into an urban area, or when traffic patterns dramatically change, we may expect a full model change and update. The rate at which model inputs or transitions are sent is expected to be much lower than the 5–10 Hz used for baseline. Model transitions usually happen at a rate much lower than 1 Hz (perhaps every few minutes), while model input updates may need higher rate of around 1–10 Hz. With these considerations, we layout the following rules for communication of models and updates (some typical parameter values are provided):

1. Full model construct should be communicated at least every T_m seconds (e.g., $T_m = 30$ s, m subscript refers to model), or when a change of model construct happens.
2. Model parameter or transition updates should be communicated at least every T_u seconds (e.g., $T_u = 10$ s, u subscript refers to update), unless a change in parameters or state transition that cannot be estimated is detected.
3. Model inputs should be communicated at least every T_i (e.g., $T_i = 1$ s, i subscript refers to input), or when estimation error using last input parameters is found to be high.

The overall communication strategy, considering the error-dependent approach, is depicted in Fig. 9. This framework applies to any model and modeling approach; however, the difference operator in $H(t) - \tilde{H}(t)$ needs to be defined for each type of modeling framework. $H(t)$ denotes the model and state of the sender at time t, and $\tilde{H}(t)$ is the estimated model and state generated locally at the sender. Together with the difference operator, the error/cost calculation module in Fig. 9 is responsible for quantifying the level of error in estimation that may result if a new message is not sent (and estimation continues with previous model and data). This module implements a cost function C of the difference, which is provided to the communication logic. Communication logic is the component that implements the rules described for each type of the modeling scheme (e.g., the three rules above).

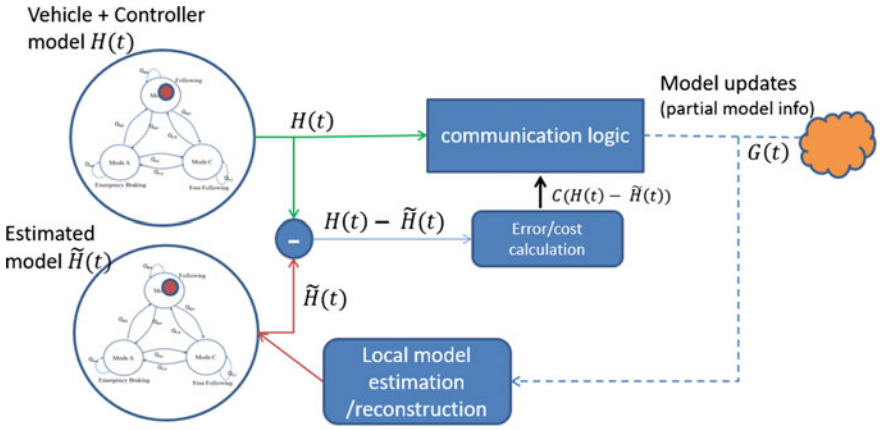


Fig. 9 General framework of error-dependent communication logic with MBC

Given the model presented in Fig. 7, and taking the example of Fig. 8, we can use the above rules and set up an experiment to see the gains made by MBC. For a clearer observation, we need to consider an application and a specific model form. The next subsection describes, in a more formal way, how a hybrid automaton describing vehicle movement can be presented. Later, we take an application such as collision warning and evaluate MBC performance.

5 Using a Hybrid Automaton Model in MBC

The example here follows a hybrid automata concept, which is a well-known method for mixed discrete-continuous state system modeling. The model can be formally described as follows:

$$H = (Q, X, Init, \Sigma, f, Dom, E, G) \tag{4}$$

where Q is a finite set of discrete states (or modes A, B, C as in Fig. 7). X is a set of continuous states and represents the equations for each mode as in (1), (2) and (3). Any pair of $(q, x) \in Q \times X$ is referred as the state of the hybrid automaton. $\sum$ is a set of discrete input symbols and events. $Init \subseteq Q \times X$ is the set of initial values of the system $(\hat{X}_j, \tilde{X}(0))$. $f(\ldots) : Q \times X \rightarrow p(X)$ represents the dynamic of the system. $Dom : Q \rightarrow p(X) \times \Sigma$ is the working domain of each state and defines combinations of states, events, and constraints for which dynamical equations are allowed; here $p(x)$ is the power set of all subsets of X. $E \subseteq S \times S$ is the set of edges, and $G : E \times \Sigma \rightarrow p(X)$ is the set of transition guard conditions that enable the transition between discrete states.

The form in (4) can be encoded using standardized notations such as XML; the size of the construct is expected to be small and less than 500–1000 bytes. This form can usually be communicated in a single message. Further model updates and inputs can use even smaller packets that only inform receivers of updated input values or of a transition between modes. For the example of Fig. 7, the values of V_{LV}, and a_{LV} are the input updates that will be communicated at a rate R. Noting Fig. 8, the LV is the vehicle in front of the sender of the model which is shown as RV. RV may have this information either through its own estimated model of the LV (LV may also use MBC to send info to RV); it is also possible that the receiver (HV) directly hears the position information that LV broadcasts through BSM messages of baseline or MBC communication.

It is noteworthy that the hybrid automaton form described above is only an example and other model forms can be used with MBC strategy. Later in this section, we introduce an adaptive model form that can be constructed and updated using online machine learning methods.

6 Evaluating MBC with Collision Warning Application

We next evaluate the utility of MBC in a cooperative collision warning application. The sample application is a cooperative version of the Forward collision warning (FCW) application which produces a warning to the driver if the threat of collision with the car in front is detected. The crash detection method of FCW can be adjusted for automated crash avoidance as well. In FCW, a given vehicle (called a Host Vehicle or HV) continuously analyzes the position and movement of its surrounding vehicles to detect the possibility of a crash with other vehicles. The analysis happens on a real-time map that is built using information received from other vehicles (called remote vehicles or RVs as shown in Fig. 8). In an FCW application, the vehicle immediately in front is the main subject of analysis for threat detection.

Traditional (noncooperative) FCW systems use ranging sensors (e.g., lidar or radar) to detect others vehicles in front. In cooperative FCW, information available through wireless communication (for example over DRSC) can be used to detect the presence and position of other vehicles. Apart from forming a real-time map and detecting other vehicles, the cooperative or noncooperative versions of FCW can be the same. In this chapter, we look at a publically available algorithm for FCW, called CAMPLinear [19] and assume that information from other vehicles is available in this algorithm over a vehicular network.

Collision hazard detection in CAMPLinear is done by comparing the distance between HV and RV (d_s) with a variable warning range (r_w). If the distance becomes less than the warning range r_w, a hazard situation is assumed (detected) and an action needs to be taken (see Fig. 10 for an example of typical values). The main component of CAMPLinear is the algorithm that produces a warning range value based on the current situation and movement of HV and RV. Both the warning range and the distance between HV and RV change over time and need to be continuously

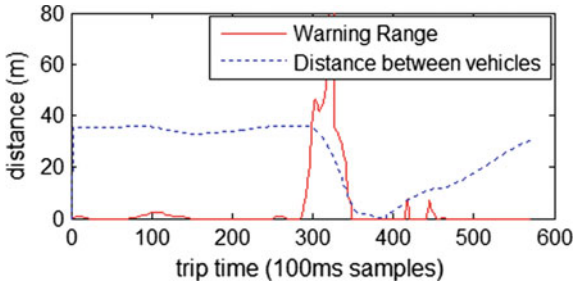


Fig. 10 Example of warning range derived from CAMPLinear algorithm, produced on data from a near crash scenario from 100-car dataset

calculated and monitored. Details of the CAMPLinear algorithm can be found in [19]. While we do not repeat the details here, a summary of how this algorithm works is described below.

6.1 A Sample FCW Algorithm

The CAMPLinear algorithm calculates the warning range r_w using speed and acceleration data of RV and HV (i.e., $v_{RV}, a_{RV}, v_{HV}, a_{HV}$). The r_w value is computed as the sum of a brake onset range (*BOR*) parameter and the distance traveled by the vehicle during reaction time of the driver/system (denoted as t_d). Therefore, we can write r_w as

$$r_w = BOR + (v_{HV} - v_{RV})t_d + 0.5(a_{HV} - a_{RV})t_d^2 \tag{5}$$

The value of t_d (driver and brake system reaction delay) is found to be around ~2.5 s. BOR is computed for several different scenarios for stationary or moving RV. For the sake of the discussion in this chapter, we only show BOR for the moving RV as follows; other cases are described in detail in [19]:

$$BOR = (v_{HVP} - v_{RVP})^2 / \left(-2\left(a_{rqd} - a_{RV}\right)\right) \tag{6}$$

where v_{HVP} and v_{RVP} are the predicted speeds of HV and RV after a t_d time. The assumption is that we have the values at the warning moment, so the predicted values are for t_d second in the future are found as: $v_{HVP} = v_{HV} + a_{HV}.t_d$ and $v_{RVP} = v_{RV} + a_{RV}.t_d$. a_{RV}. The value of a_{RV} is the acceleration of RV; a_{rqd} in Eq. (6) is the acceleration (deceleration) required at the HV for avoiding a crash and is modeled in [19] using actual human reaction data, as follows:

$$a_{rqd} = -5.3 + 0.68a_{RV} + 2.57U(v_{RV}) - 0.086(v_{HV} - v_{RVP}) \tag{7}$$

where $U(v_{RV})$ represents the unit step function that results in 1 when the speed of RV is positive, and 0 otherwise. We are using a slightly different notation in Eqs. (5)–(7), than what is used in [19].

A vehicle that implements CAMPLinear FCW algorithm needs to continuously (at regular intervals) run the algorithm to watch for possible hazards. The execution interval is usually set to 100 ms. The information used by the FCW algorithm comes from several sources. It is assumed that a vehicle has precise information about its own movement and position (thus, information about HV position, speed, and acceleration is assumed to be correct). However, RV movement information is only available as an estimate for the above calculations. Since data about RV is received over a DSRC network, an estimation process is usually utilized to produce more reliable information on RV movement. In the baseline design, a constant speed (or constant acceleration) model may be used for estimation of RV position using data received over the network. With MBC, it is possible to use more complex or dynamically changing models. For the example studied in this chapter, we assume that RV movement can be modeled using a 3-mode car following model of Fig. 7. With this model, and the communication rules described above, we can evaluate how the FCW algorithm performs.

6.2 Performance Metrics

For performance evaluation, we first need to define a few **metrics**. Given the structure of the FCW system which utilizes an estimation process followed by a hazard detection algorithm, we consider two metrics of estimation error for movement data, and accuracy of crash detection (or alert generation). The estimation error can be considered an intermediate measure since its impact is only indirectly seen on the FCW algorithm accuracy (which is directly observed by a user of the system). For the estimation metric, we focus on position tracking error which is often used in the study of DSRC-based CCW systems [4, 1]. For accuracy of the FCW algorithm, we consider an "alert accuracy" or "hazard detection accuracy" measure where the accuracy of hazard detection in calculations that happen at 100 ms intervals is evaluated versus a ground truth situation. The ground truth is established by assuming that HV has complete information about RV movement.

6.3 Communication Strategies

To evaluate the performance of MBC, we consider three communication strategies of (1) baseline, (2) MBC-1 with model input updates only, and (3) MBC-2 with model input updates and transition updates. The baseline method is assumed to use constant acceleration model which is more accurate than constant speed model for the dataset used in our simulations. Both MBC options assume that each vehicle sends out its

movement model once at the beginning of the test (rule 1), and then updates to the inputs of the model are periodically transmitted. With MBC-2 settings, all messages of MBC-1 are transmitted, in addition to occasional messages when the driving mode transitions happen (e.g., changing from following to emergency braking mode in Fig. 7); such messages are very rare.

The main difference between MBC-1 and MBC-2 is that mode changes are immediately included in communications in MBC2. While it is possible that even with MBC-1, the receiver estimates a mode change, the estimation may be delayed or not estimated if the guard conditions do not depend on the state refinements. When mode change is due to change in input values that are also included in model updates and sent to the receiver, we expect the receiver to deduce the mode change from its available data (perhaps with some delay). But when mode changes happen due to other external inputs that may not be immediately available at the receiver (like a driver's visual input, or probabilistic mode changes), it is not expected that the receiver estimates such transitions. If state transitions are included in regular model updates, there will be a delay until they are detected at the receivers. In MBC-2, we assume that a transition is assumed to be a large error in the error-dependent logic; as a result, a transition causes an immediate message transmission to the receivers, minimizing receiver's delay in tracking sender's current state.

7 Evaluation Using 100-Car Naturalistic Driving Data and Car Following Models

Using the 100-car naturalistic driving data [20], a simulation study is set up as in Fig. 8. Assuming a single lane, the first vehicle in the lane is moved according to the trajectory of a vehicle from the 100-car dataset. There are over 800 scenarios of near crash or crash recorded in this database. Movements of other vehicles in the scenario of Fig. 8 used the 3-mode car following model from [11], also depicted in Fig. 7. The simulation platform is a modified version of our earlier work in [8] and depicted in Fig. 11.

The first performance metric, speed estimation error, is plotted in Fig. 12. To clearly demonstrate the results and for a fair comparison, we set the message rate for all three methods of baseline, MBC-1 and MBC-2 at a low rate of 2 Hz (with no loss). It must be noted that a rate of 2–3 Hz is what is generally generated using a practical method such as ED [1]. It is observed that MBC-1 and MBC-2 have significantly lower number of large speed estimation error instances than baseline. With the use of MBC-2, all large errors instances are avoided by communicating mode transition information. The instances when speed estimation error is large appear to be around points in time that the speed of vehicle changes significantly and at a high rate. Such sudden changes are the result of drivers suddenly taking a different action like hard braking. Transition between some other modes such as from free following

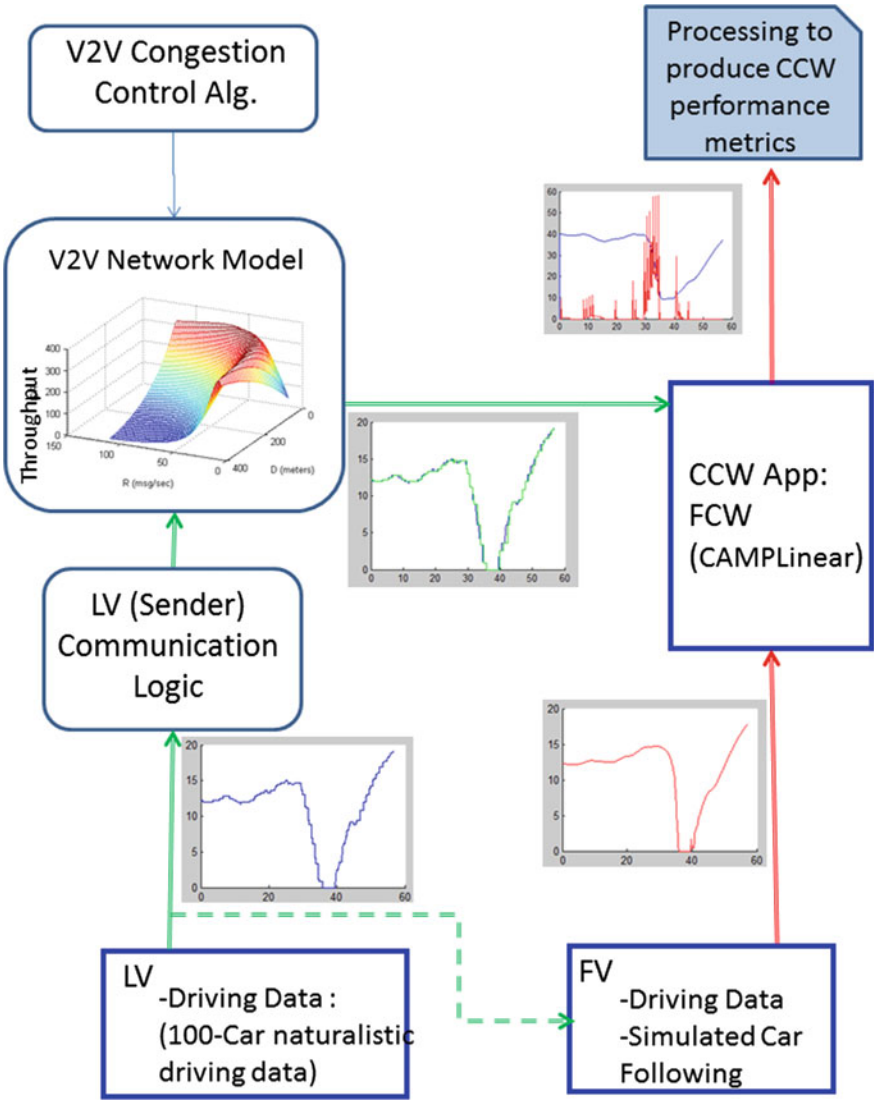


Fig. 11 CCW evaluation and modeling framework

to following does not cause large changes. In those instances, MBC-1 and MBC-2 behave more or less similarly.

Error in estimation (of position, speed, etc.) is an intermediate performance metric, since these estimated quantities are input to another component which is the FCW algorithm. To evaluate the impact of this error on FCW performance, we look at a measure called hazard detection accuracy (defined using a method similar to what

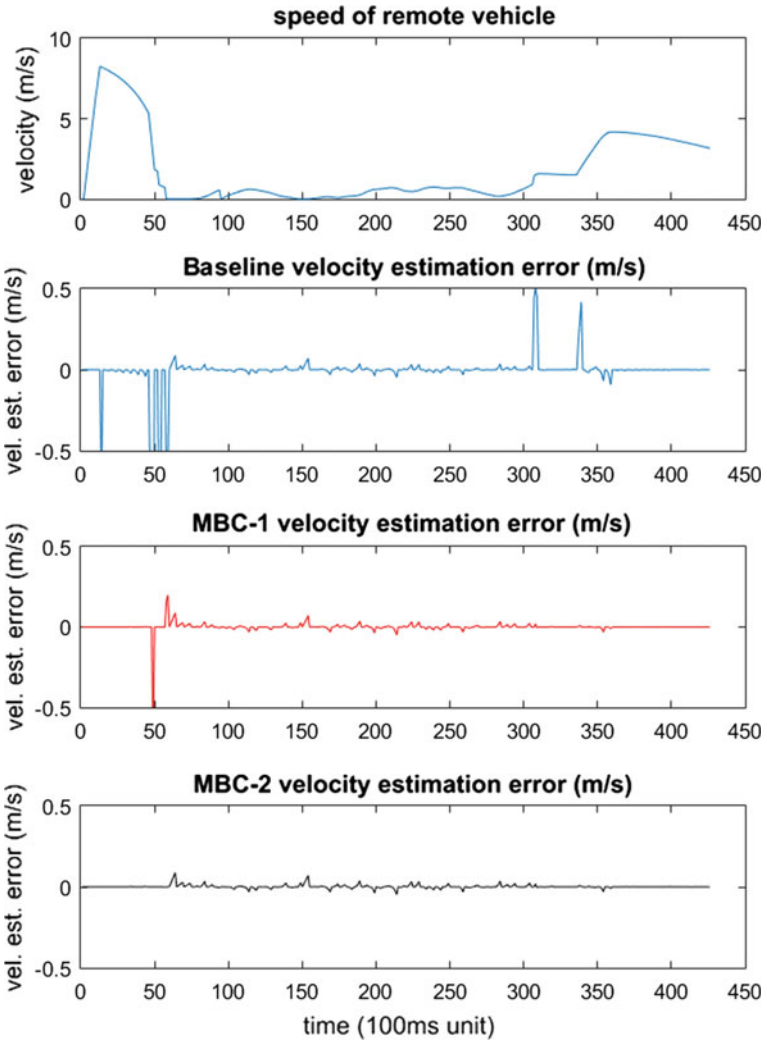


Fig. 12 Estimating speed of a remote vehicle using baseline and MBC methods

[21] presents). This measure, informally called "accuracy" in the rest of this chapter, is calculated as the ratio of instances of correct identification of the situation as hazardous (*Ch*) or safe (*Cs*) to all FCW execution instances for a scenario (*M*): $accuracy = (Cs + Ch)/M$. Note that correct identification of the situation is with respect to the ground truth results obtained from running CAMPLinear with full knowledge of the movement of RV. Since movement information in real cases of the three communication methods are only available as estimates, we expect that accuracy

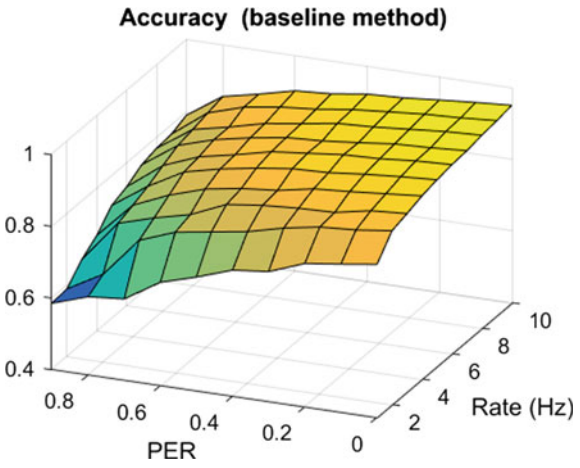


Fig. 13 Accuracy versus PER and rate for baseline method

to be always lower than 1. Wrong classifications of hazards are cases when false positives or false negatives are produced by the FCW algorithm.

Here, to see the full impact of communication, and for a fair comparison, all three methods are assumed to use the same communication rate of R Hz and the same packet lengths for the messages. With MBC methods, occasional model and transition information are communicated which in the scenarios studied here will amount to less than 2 or 3 instances of communication in an average of 300–500 messages. Adding this negligible extra communication, which is less than 1%, to the baseline method will not improve the baseline results in an observable way. All different choices of rate are also tested under various network PER values between 10 and 90%.

Results are shown in Fig. 13 for baseline, and Fig. 14 for MBC. In these figures, hazard detection *accuracy* is plotted versus all choices of *Rate* of information transmission (from sender) and *PER* of the network. From these figures, it is obvious that MBC has a significantly better accuracy at high PER and low rates. These are the situations in which information at the receiver is available at a low rate due to congestion or high packet loss [1]. As expected, in low PER and high rate situations, the two methods perform well.

It is important to note that the hazard detection accuracy values should not be interpreted in an absolute sense. Since accuracy is calculated as a ratio of $(Cs + Ch)/M$, and moments when estimations are wrong are rare, the accuracy numbers are usually high even when critical situations are missed. To put this in perspective, consider the results in Fig. 12, for a 42-second trip. It is seen that for most of the trip time, estimation error is very low, and only for a small number of instances, the error becomes high. In safety applications, these moments are of high importance, however their impact on the numeric value of accuracy will be low. Also, if the same scenario of Fig. 12 continues by another 42 s but without sudden speed change (which is normal in most of the driving situations), the difference in accuracy under different

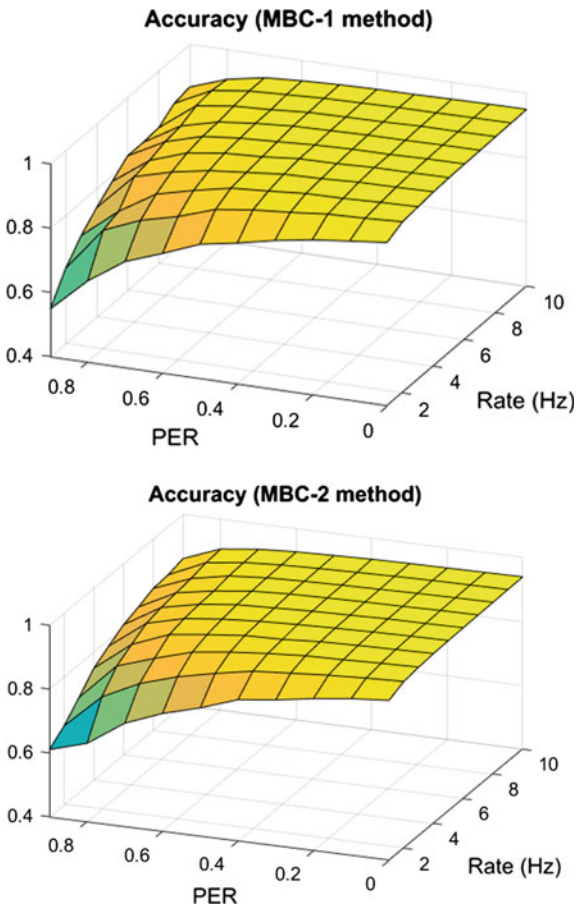


Fig. 14 Accuracy versus PER and rate for MBC approaches

situations becomes half. Therefore, we should see accuracy values in a relative sense. In general, safety critical events in driving are rare events, so accuracy values should be high to ensure that these rare but important moments are covered as much as possible.

To see the difference between baseline and MBC results better, we plot the results for these methods in one figure for two situations. The first situation, for PER of 0 and rates varying from 1 to 10, is shown in Fig. 15. In Fig. 16, the rate is fixed at 10 Hz, but PER values are varied from 0 to 0.9. It is observed that in both settings, MBC is able to maintain a high value of accuracy, above 90%, when the rate of information received at the receiver is as low as 1 Hz. As discussed above, the difference between the accuracy of 80 and 90% can be very significant as safety events are rare and 80% accuracy may lead to unacceptable delay, missing of alerts, or too many false positives. The baseline method can reach the same accuracy of 90% or above at rates higher than 4 Hz. In addition, the accuracy of above 80% can be maintained

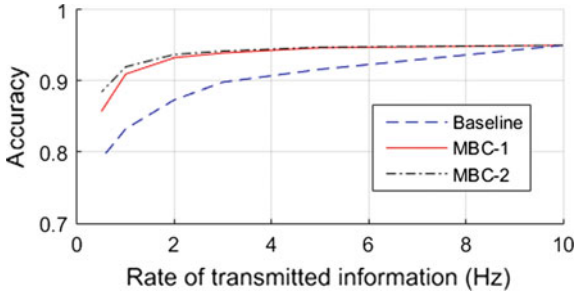


Fig. 15 Accuracy versus rate for baseline and MBC methods, assuming PER = 0

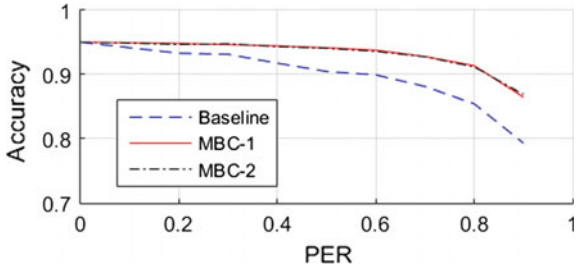


Fig. 16 Accuracy versus PER for baseline and MBC methods, (R = 10)

at PER levels of up to 0.8 PER for MBC (this is equivalent of reception rate of 2 Hz), while baseline requires PER less than 0.4 (or reception rate of 6 Hz) for the same performance. Another interesting observation in these two figures is the visible improvement that is achieved by MBC-2 over MBC-1 at low rates and when losses are not random (Fig. 15).

An interesting observation from Figs. 15 and 16 is that with the same rate of message reception at around 1 Hz, MBC-2 performs notably better when the reduction in rate is not random (in Fig. 15). In Fig. 16, where the rate reduction to 1 Hz is due to random loss in the network, the results of MBC-1 and MBC-2 are almost the same. Note that the rate of information received at a receiver can be calculated on average as *R*(1-PER)*. The reason for this difference may be due to the fact that with very high random loss at 0.9, it is very likely that mode transition updates are also lost. This points to the fact that error-dependent policy needs to be aware of network situation (as in [1] and investigated in detail in [22]).

In addition, it can be observed from these figures that a given rate of reception achieves a higher accuracy when PER is 0 and rate is reduced, compared to when the rate is 10 Hz and PER increases. The reason for this difference is that irregular packet losses (which are random) tend to create higher possibilities of miscalculating warning situations, for example, when consecutive losses occur.

While MBC is consistently outperforming baseline, the amount of improvement is different in the two situations in Figs. 15 and 16. Notably, it is seen that while accuracy of 90% is achieved at rates 1 and 4 Hz for MBC and baseline, the same accuracy requires 2 Hz (for MBC) and 6 Hz (for baseline) when sending rate is fixed at 10 and PER varies. The reduced difference between these methods is perhaps

due to the fact that in MBC, each packet has more importance and is carrying more valuable information compared to baseline (since MBC messages contain more sensitive data). This means that at higher random loss, performance degradation will be higher. Nevertheless, the MBC method needs much less data at the receiver and is far more efficient in producing accurate results when the network is under constraints. In practice, high network losses are possible in many traffic situations [1, 22]; thus, a sending rate of 2–5 Hz is generally more acceptable than 10 Hz due to congestion/scalability issues that exist in vehicular networks.

7.1 Improving MBC-Based Method with Network Awareness

An additional improvement to the MBC-based methods can come from network awareness. The network awareness feature was introduced in [1] for the error-dependent policy in DSRC vehicle safety networks, and studied in detail in [22]. The main concept behind this feature was the employment of a network loss simulator which would increase the error of the local estimator and lead to higher transmission rate when PER was high. The same method cannot be directly employed with error-dependent version of MBC, since there are multiple types of messages with different levels of importance in MBC. Nevertheless, to show that the concept can still be useful, we designed a simple addition to MBC-2 method to counter the effect of high packet loss on important messages such as transition update messages. This simple method uses the PER value as the failure probability and through a Bernoulli trial guesses whether the transition update has been successfully received. If the sender guesses that the transition update was not successful, it tries to retransmit the update up to K (e.g., $K = 3$ here) times. We call this method MBC-2 + N.A.R (network-aware redundancy); the plots in Fig. 17 show the resulting improvement. It is noteworthy that the retransmissions are expected to increase the rate and communication load; however, due to the rarity of transition updates, the increase in rate is negligible and very small. It is seen that the accuracy is increased for values of PER greater than 0.5. Since this redundancy method may not be the best option for protecting the high-value messages, we will not include it in the rest of the results in this section. Further development of the network awareness feature is a topic of future research.

7.2 Improvements to Tracking Accuracy

The accuracy improvement that MBC provides (as seen in Figs. 15 and 16) is the result of its more accurate tracking of a remote vehicle. Figure 12 showed a specific example of how MBC reduces the estimation error. To do a more comprehensive comparison, we recorded the estimation errors for the same scenarios of the above experiment (with different PER and rates) and plotted the results in Fig. 18. The error shown in this figure is an average measure computed as follows. First, the 95

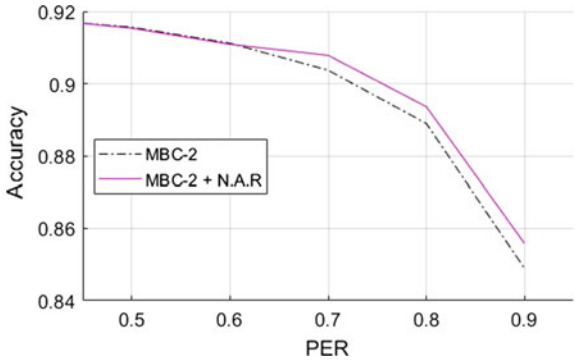


Fig. 17 Improving accuracy at high PER values through network-aware redundancy feature in MBC-2+N.A.R

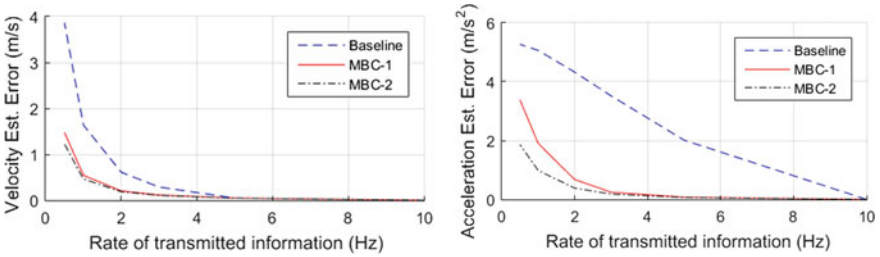


Fig. 18 Velocity and acceleration estimation error (95 percentile) for different communication methods

percentile of error in each scenario of the simulation (built using 100-car dataset) was recorded, and then the mean of all 95 percentile errors was taken as the estimation error. We took the 95 percentile error since it is routinely used for evaluation of position tracking error for safety applications [1]. The results show the same trend as in Fig. 16. An interesting observation is that the improvement in estimation error is more visible for acceleration estimation. We expect this to be due to the fact that the 3-mode movement models produce acceleration values directly, while speed is an integration of acceleration and the improvements are more indirect. Overall, given the reliance of FCW algorithm on both acceleration and speed values, the improvements are easy to note in measures like accuracy.

8 An Example of MBC with Model Construct Changes

The movement model example presented above, and used with MBC, assumed that the model construct remains valid for the entire duration of the system run time (in this case simulation time). In more realistic situations, where traffic conditions and environment of a car changes significantly, it may be necessary to use a different

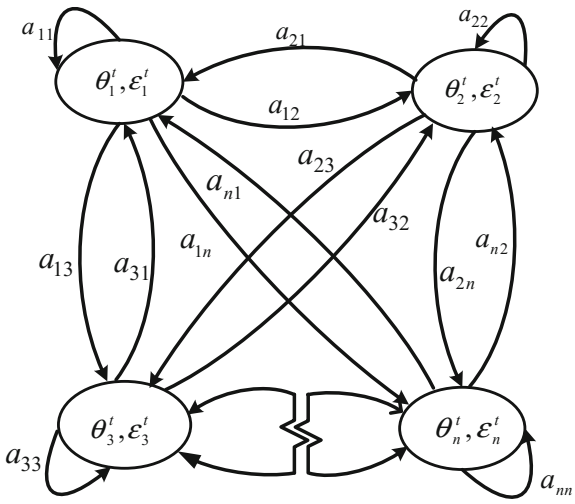


Fig. 19 An adaptive SHS with variable number of modes, for modeling vehicle acceleration profile

model. In such cases, MBC requires an update to the model construct. In fact, we expect MBC to demonstrate even more significant gains in such situations.

An example of models that change over time is presented in [10]. The approach in this chapter is to use a stochastic hybrid system (SHS) form that comprises of a hidden Markov model (HMM) with modes that use AutoregRessive eXogenous (ARX) equations as state refinement. The HMM model is assumed to evolve over time and is evaluated on a regular basis to determine whether new modes need to be added to the model. The work in [10] assumes that with MBC, the adaptive model will allow for the best ARX representation to be available at any given time. The ARX formulation is used for modeling acceleration changes and for control purposes in a model predictive controller. In this chapter, we demonstrate how MBC can use the model in [10]; details of communication logic rules for updates on model and model constructs are elaborated using this example in the rest of this chapter.

The model presented in [10] is shown in Fig. 19. In this HMM, transition probabilities are shown as a_{ij}, and each mode has parameters that describe the state refinement as an ARX model for the acceleration (probabilities for output generation and ARX parameters). Parameters θ_{ij} are the regressor coefficients of the ARX model and $\varepsilon(t)$ is the ARX approximation error, and is expected to have a zero mean Gaussian distribution. While how the model in FIG is derived and trained is out of the scope of this chapter, we need to specify how such a model can be communicated in MBC.

The changes that happen in this model are expected to be in the form of addition (or removing) of modes. New parameters for an ARX model are usually accounted for as a new mode in the adaptive SHS (HMM), and the parameters of existing modes are kept intact. As a result, tracking this model requires knowing the set of modes and their parameters (the model construct) as well as the current active mode. Since transitions are probabilistic, the next transitions can be estimated based on probabilities; nevertheless, transition updates can reduce the possibility of error in

that estimation. When a new mode is added, the estimation error at a receiver that is not aware of the new mode is naturally expected to be higher. Therefore, mode additions are considered to be important events that require faster communication and update.

8.1 Rules for Communication of Models

Assuming that a local modeler in the sender is maintaining a real-time accurate model of the host vehicle, the communication logic for this model can be described as follows:

- If there is an update to model construct, transmit a message containing the new model
- If the current most likely mode (from HMM) is different from the previous most likely mode, due to change in inputs, transmit the inputs and indicate the current most likely mode.
- If model inputs (to ARX) do not result in change of mode but result in error greater than a predefined threshold, update the receiver with the current inputs.

Note that the advantage of using the probabilistic multimodal form above is that the mode changes can be estimated in many cases without the need for explicit communication. The above rules, when applied to the ARX-HMM models, allow the receiver to not only track the movement of the sender but also provides a model (i.e., ARX in this case) which can be used for other purposes such as model predictive control. The work in [10] uses the ARX models in an efficient MPC-based cooperative adaptive cruise control method. As the details of such designs are outside the scope of this chapter, we do not further repeat them here. However, a glimpse into the results reported in [10] shows that the advantages of MBC are twofold. First, communicated models can help produce more accurate estimations or reduce the communication load. Second, models available at the receivers can be used in model predictive control for automated applications.

9 Concluding Remarks

Communication in cyber-physical systems is traditionally viewed from the perspective of sensing a physical or cyber process and reconstructing sensed signals at the receiver side. Models of the physical or cyber phenomena are used at the receiver to help reconstruct (or estimate) the original signal or recover from communication losses through estimation. Recently several methods (such as the error-dependent communication policy) have been proposed that utilize the models in the sender, in addition to the receiver, to make communication events more efficient and relevant

to signal variations. This chapter describes a transformation of the use of models to a more general concept of Model Communication.

The concept of model-based communication transforms communication in CPS to a model-based paradigm, where models and model updates replace raw sensed data. This concept is beyond using models for the purpose of reconstructing sensed signal, and aims also at communicating and reconstructing models (which in turn produce signals). Communicated messages may include complete model constructs, updates to model states, updates to model parameters, transition updates for multiple state models, or simple data that is used as input to models.

It is shown that the concept of MBC can result in significant improvement in accuracy of estimation, or equally in reduction of the communication load needed for estimation. From a high-level perspective, the improvement is due to the sender and receiver having better knowledge of the signal with the use of models. More specifically, the continuous update of models ensures that the physical or cyber phenomenon is better tracked over time. Another advantage of MBC is the possibility of receivers using current and real-time models of senders in model predictive control applications. The compound effect of these two improvements (improved estimation, and model predictive control) is shown in some of our recent papers, while this chapter elaborates on the MBC logic and the basic concepts behind MBC.

Utilizing the concept of MBC requires determining models and rules for communication. The specific rules and algorithms governing communication under MBC depend on the modeling approach and frameworks chosen for a given application. This chapter described general rules for error-dependent communication under MBC; specific rules for two example models were discussed in more detail. We took the example of vehicle movement, and applications such as collision warning to demonstrate how MBC can be used in practice. While significant improvements were shown in this example, we also note that the choice of models and communication rules are very important and will have a significant impact on the outcome. In fact, the challenges of determining modeling schemes, models and communication rules are important challenges that need to be addressed when MBC is used. We expect a formulated set of rules for general stochastic hybrid systems forms which will be applicable to the above applications as well as many other CPS-related applications. Such mechanisms are still under study.

Given that SHS models can be general enough for many CPS applications, a natural next step in evolving the concept of MBC is to develop MBC rules specific to SHS forms. It must be noted that models for specific phenomena do not need to be predetermined or fixed. It is possible to learn models and update them even at the sender side in real time (as demonstrated in [10]), the changes are then communicated to receivers following MBC rules. This possibility further underscores the potential of using MBC for highly dynamic systems.

References

1. Huang, C.-L., Fallah, Y.P., Sengupta, R., Krishnan, H.: Adaptive intervehicle communication control for cooperative safety systems. IEEE Network **24**(1), 6–13 (2010)
2. Xu, Y., Hespanha, J.: Estimation under uncontrolled and controlled communication in networked control systems. In: Proceedings of Conference on Decision and Control, Dec. 2005
3. Vehicle Safety Communications—Applications (VSC-A) Final Report, Technical Report DOT HS 811 492A, September 2011
4. Rezaei, S., Sengupta, R., Krishnan, H., Guan, X., Bhatia, R.: Tracking the position of neighboring vehicles using wireless communications. Elsevier J. Transp. Res. Part C Emerg. Technol. SI Veh. Commun. Netw. **18**(3), 335–350 (2010)
5. IEEE 802.11 WG, Part 11: Wireless LAN Medium Access Control (MAC) and Physical Layer (PHY) Specifications, IEEE (2007)
6. SAE J2735 standard, Dedicated Short Range Communications (DSRC) Message Set Dictionary, March 2016
7. SAE J2945/1 standard, On-Board System Requirements for V2V Safety Communications (2017)
8. Fallah, Y.P., Khandani, M.K.: Analysis of the coupling of communication network and safety application in cooperative collision warning systems. In: Proceedings of the ACM/IEEE Sixth International Conference on Cyber-Physical Systems. ACM (2015)
9. Fallah, Y.P.: A model-based communication approach for distributed and connected vehicle safety systems. In: Proceedings of the IEEE Systems Conference (2016)
10. Moradi-Pari, E., Mahjoub, H.N., Kazemi, H., Fallah, Y.P.: Utilizing model-based communication and control for cooperative automated vehicle applications. IEEE Trans. Intell. Veh. (2017)
11. Yang, Q.: A Simulation Laboratory for Evaluation of Dynamic Traffic Management Systems. Department of Civil and Environmental Engineering, Massachusetts Institute of Technology, vol. 193 (1997)
12. Sekizawa, S., Inagaki, S., Suzuki, T., Hayakawa, S., Tsuchida, N., Tsuda, T., Fujinami, H.: Modeling and recognition of driving behavior based on stochastic switched ARX model. IEEE Trans. Intell. Transp. Syst. **8**(4), 593–606 (2007)
13. Kim, J.H., Hayakawa, S., Suzuki, T., Hayashi, K., Okuma, S., Tsuchida, N., Shimizu, M., Kido, S.: Modeling of driver's collision avoidance maneuver based on controller switching model. IEEE Trans. Syst. Man Cybern. B Cybern. **35**(6), 1131–1143 (2005)
14. Okuda, H., Ikami, N., Suzuki, T., Tazaki, Y., Takeda, K.: Modeling and analysis of driving behavior based on a probability-weighted ARX model. In: IEEE Trans. Intell. Transp. Syst. **14**(1), 98–112 (2013)
15. Alur, R., Dill, D.L.: A theory of timed automata. Theoret. Comput. Sci. **126**(2), 183–235 (1994)
16. Lynch, N., Segala, R., Vaandrager, F.: Hybrid i/o automata. Inf. Comput. **185**(1), 105–157 (2003)
17. Hespanha, J.: Modeling and analysis of networked control systems using stochastic hybrid systems. IFAC Ann. Rev. Control **38**(2), 155–170 (2014)
18. Hu, J., Lygeros, J., Sastry, S.: Towards a theory of stochastic hybrid systems. In: Hybrid Systems: Computation and Control. LNCS 1790, pp. 160–173. Springer, Heidelberg, Germany (2000)
19. Kiefer, R., Cassar, M.T., Flannagan, C.A., LeBlanc, D.J., Palmer, M.D., Deering, R.K., Shulman, M.A.: Forward collision warning requirements project: refining the CAMP crash alert timing approach by examining 'last-second' braking and lane change maneuvers under various kinematic conditions. Report No. DOT-HS-809-574. Washington, DC: National Highway Traffic Safety Administration (2003)
20. Dingus, T., Klauer, S.G., Neale, V.L., Peterson, A., Lee, S.E., Sudweeks, J., Perez, M.A., Hankey, J., Ramsey, D., Gupta, S., Bucher, C., Doerzaph, Z.R., Jarmeland, J., Knipling, R.R.: The 100-Car Naturalistic Driving Study, Phase II—Results of the 100-Car Field Experiment. National Highway Traffic Safety Administration, Washington, DC (2006)

21. Lee, K., Peng, H.: Evaluation of automotive forward collision warning and collision avoidance algorithms. Veh. Syst. Dyn. **43**(10) (2005)
22. Fallah, Y.P., Khandani, M.K.: Context and network aware communication for connected vehicle safety applications. In: IEEE Intelligent Transportation Systems Magazine (2016)

Part II
Internet of Things

Urban Microclimate Monitoring Using IoT-Based Architecture

M. Jha, A. Tsoupos, P. Marpu, P. Armstrong and A. Afshari

Abstract In this chapter, various aspects related to Internet of Things (IoT) based sensor node development for urban microclimate monitoring are presented. The discussion is focused on software development, relevant methodologies, hardware modules, and platforms. A typical sensor node consists of sensors, computing/controlling unit, and a communication unit. There is a large variety of environmental sensors (temperature, wind, humidity, etc.), computing units (single-board computers and microcontrollers), and communication units. With the rise of Internet-enabled devices, the ideas of IoT are being incorporated into sensor node development.

1 Introduction

Climate can be defined as a set of specific characteristics or a pattern of weather variables such as temperature, humidity, wind, precipitation, pressure, etc., of a specific region. Microclimate is the distinctive climatic characteristic of a relatively small area or locality, such as a park, garden, or a particular section of a city. In an urban setting, the microclimate is affected by various local conditions such as traffic, building architecture, shading, presence of a waterbody, etc. In a microclimate, the weather variables may vary from the ones that prevail in adjacent areas at a given instant of time due to the difference in geographical/physical conditions and

M. Jha · A. Tsoupos · P. Marpu (✉) · P. Armstrong · A. Afshari
Masdar Institute, Abu Dhabi, UAE
e-mail: pmarpu@masdar.ac.ae

M. Jha
e-mail: mjha@masdar.ac.ae

A. Tsoupos
e-mail: atsoupos@masdar.ac.ae

P. Armstrong
e-mail: parmstrong@masdar.ac.ae

A. Afshari
e-mail: aafshari@masdar.ac.ae

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_4

anthropogenic activities. The different microclimatic conditions in various areas/localities inside a city constitute the city's microclimate, i.e., urban microclimate.

According to the United Nations, the urban population will grow up to 6.3 billion and about 67% of the world's total population will be living in urban areas by 2050 [24]. As of today, the urban areas occupy only 2% of the earth landmass, consume about 75% of the world's energy and produce 80% of total greenhouse gases [36]. When large areas are urbanized, urban climate is affected significantly creating diverse microclimatic regions within the same city. Hence, a methodological approach should be adopted to systematically study the urban microclimate to understand the land use and land cover interactions with the local microclimates.

Human activities, infrastructure, landscape, and pollution have a major influence on urban microclimate. In urban areas, weather variables like wind, temperature and humidity, etc., are affected by high-rise buildings, shading, the road network, heat exhausted from factories, air-conditioners, topography, and vegetation. The building material also plays a crucial role in the urban microclimate patterns.

There are various examples of microclimates such as greenhouse, heat island, inverse heat island effects, etc. Microclimates can be generated under various conditions such as presence of waterbody which results in lowering the temperature compared to surrounding, presence of slopes and contours which result in varying sunshine on the landscape, presence of anthropogenic heat, and presence of vegetation.

Urban microclimate is very dynamic because of the number of factors involved and the associated complexity of interaction among those factors. Hence, the first step to study the urban microclimate is to measure several climatic variables and study the correlations. The urban microclimate data will be further useful to validate simulation-based urban microclimate models being developed by various researchers [37, 60, 62]. To study urban microclimate, a network of sensors needs to be installed throughout the urban grid and acquire the sensor data. In traditional sensor network paradigm, the data is acquired from the sensor nodes by physically accessing the data logger or by routing data to a data collection center by intranet methods such as ZigBee, Ad Hoc networks, etc. However, with the IoT paradigm, as the sensors are directly connected as things within a broader network, the data should be accessible instantaneously.

The study of microclimate is important because the insight generated from microclimate study will enable policymakers, architects, and designers to create urban infrastructure which will be conducive for human comfort. Urbanization process has a significant impact on local climate [29, 30]. Study of urban microclimate will facilitate various decision-making processes such as power demand, outdoor thermal comfort, pollution status, etc., for the city planners [30, 37]. In dynamic urban environments, the microclimate affects the quality of life. The other important application of urban microclimate weather data is validation for urban microclimate models proposed by various researchers [37, 60, 62]. One popular approach to study urban microclimate is the use of satellite data. The problem with satellite data is low spatial and temporal resolution. In situ sensor networks can overcome these issues by installing high temporal resolution sensor nodes with the required spatial resolution.

To enable the communication capabilities over Internet, we need to associate Internet of Things paradigm with the sensor nodes.

Although the idea of Internet of Things (IoT) was first proposed in 1999, the International Telecommunication Union (ITU) formally proposed the concept of IoT in 2005 in the "World Summit on the Information Society (WSIS)" [33]. The main principle of IoT is that "things" (i.e., sensor nodes in our case) should be able to identify, sense, process, and communicate, without human intervention [33]. IoT presents a new paradigm in the Information and Communication Technology (ICT) field. It is a multidisciplinary concept which includes a wide range of technologies, device capabilities, and application domains [20]. With the advent of IoT and corresponding enabling technologies such as low-power hardware platforms, easy-to-use APIs for application development and network integration, an increasing number of modern sensor networks are implemented based on the IoT paradigm [68]. IoT-based microclimate monitoring is feasible in urban environments due to the ubiquitous availability of telecommunication service and coverage. Wireless Sensor Networks (WSNs) and Mobile Ad Hoc Networks (MANETs) are key technologies for enabling IoT applications in cities.

Considering the dynamic aspects of urban microclimate, sensor network technologies, and the IoT concept, IoT-based urban microclimate monitoring architectures pose as flexible solutions capable of covering the application requirements. Section 2 presents the literature review on urban microclimate, IoT, and use of IoT for urban microclimate monitoring. In Sect. 3, hardware and software design considerations of IoT-based sensor nodes are discussed. Further, implementation details of an urban microclimate monitoring wireless sensor network installed in the city of Abu Dhabi are given in Sect. 4. Finally, conclusions are provided in Sect. 5.

2 Literature Review

Urban microclimate demonstrates distinct patterns related to local conditions and surroundings such as human activities, infrastructure, and geographical characteristics [13, 65]. The interaction between a variety of factors affects weather variables and results in variations and specific patterns of wind and thermal flows.

Among the most common urban microclimate phenomena is the Urban Heat Island (UHI) effect. UHI is defined as a physical phenomenon in which the main city demonstrates higher air temperature than its surrounding rural or suburban areas [64]. This effect was first described by Luke Howard in the 1810s [44]. Since then, several researchers have studied the UHI effect using satellite images and in situ measurements (Fig. 1).

A lot of researchers have utilized the Land Surface Temperature (LST) obtained from satellite imagery as a proxy to calculate the various patterns of ambient temperature [29, 50, 63]. Although LST and ambient temperature have been shown to follow similar patterns, the full dynamic profile caused by various urban heat fluxes cannot be captured through LST. In order to study the UHI effect and investigate ways

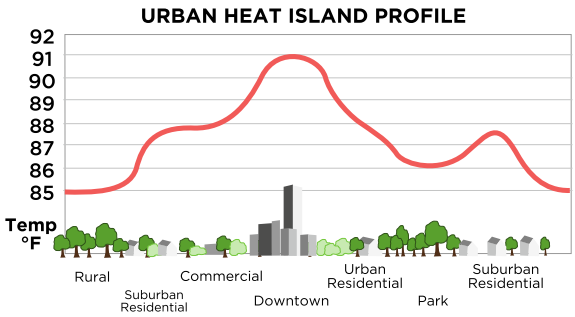


Fig. 1 Urban heat island: the temperature increases in the areas with high urbanization

of mitigating it, in situ measurements of weather variables with high temporal and spatial resolution are required. Common weather variables required for such analysis are wind speed and direction, solar irradiance, ambient temperature at different heights, etc. To perform reliable and accurate measurements of the aforementioned variables multiple specialized weather stations have to be installed in the urban grid.

In [30], Lazzarini et al. observed that the urban areas can be cooler than the surrounding rural areas in desert cities, i.e., inverse heat island effect. This observation was based on remote sensing data of land temperature. The remote sensing data source does not have the required temporal and spatial resolution to perform a fine-scale analysis of urban environment [30].

Sobrino et al. suggested that for proper study of UHI, a spatial resolution of at least 50 m is required [62]. The high-resolution satellite thermal data from LANDSAT has a resolution of 60 m [29]. Also, satellite data only gives land surface temperature and hence, temperature gradients based on height cannot be determined. Furthermore, the relation between land surface temperature and ambient temperature is complex in urban areas. These factors establish that satellite-based analysis of UHI effect has limitations. This can further be extrapolated to the requirement of a network of sensor nodes to study urban microclimate with high spatial and temporal resolution. Further, many useful applications can be developed over citywide sensor network such as dynamic waste collection, smart parking, efficient public lighting, etc.

IoT is a new paradigm in ICT advancements which is basically connecting the things (i.e., sensors node in our case) to the Internet to facilitate data acquisition, diagnose issues with sensors, and monitor the sensor node status, without active physical intervention of human [33]. Since IoT-enabled sensor node is a multidisciplinary concept which includes knowledge base from environmental, computer, electrical, and electronic science domains, it calls for a collaboration between these fields to develop successful IoT-oriented solutions.

IoT is one of the fastest growing technologies in the world as shown in the graph Fig. 2. The International Data Corporation (IDC) [34], predicted that 28.1 billion IoT devices will be installed by 2020, as shown in the graph Fig. 2.

Anagnostopoulos et al. [7] proposed an IoT-enabled dynamic waste collection and delivery to processing plants system. It is essentially an urban citywide sensor network which detects the status of waste bins and dynamically schedules the fleet of

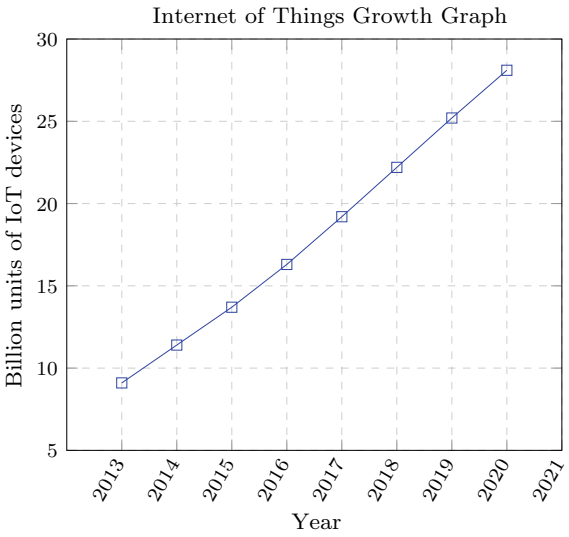


Fig. 2 IoT growth: by 2020 it is predicted that IoT devices will be around 28.1 billion

garbage truck in near real time. The rise of reliable sensors, actuators, and ubiquitous mobile communications has enabled IoT to offer dynamic solutions aimed at optimizing the garbage truck fleet size, collection routes, and prioritize waste pick-up [7]. In [35], a smart parking system based on the integration of various IoT enabling technologies, e.g., Radio-Frequency IDentification (RFID), Near Field Communication (NFC), Wireless Sensor Network (WSN), Cloud, and mobile communication, was proposed. The parking system proposed was essentially an application of a sensor network. The parking system was able to collect the environmental parameters and determine the occupancy state of parking spaces based on the sensors data input. It further directed the drivers to the nearest vacant parking spot. In [53], an easy-to-deploy low-budget IoT control system was proposed with the aim of allowing cities to cut down on electricity costs that correspond to public lighting. Its implementation was based on cellular networks and scalable Cloud computing architectures. Another application of sensor networks could be livestock management. In Australia, all the cattle are "equipped" with RFID tags as required by law [21]. Each cattle can be individually monitored, and applications can be developed for maximizing pasture utilization and preventing disease spreading by geo-fencing [21].

Various ICT projects on cities present many opportunities and challenges [42] for diverse applications in multiple domains such as decision support systems, power grids, and service-oriented architectures. These projects highlight the requirement to equip the cities with a variety of urban sensors at multiple locations to monitor the city in real time [41]. Such technologies become crucial to the emerging concept of "smart cities". The authors of [31] proposed a platform to manage urban services which includes convenience, safety, health, and comfort. To enable some of these services, a citywide microclimate monitoring system is required. Bellavista et al. [11] argued that low-cost and easily deployable WSNs and MANETs are enabling

wide-scale urban climate monitoring. The convergence of MANET and WSN enables the development of IoT platforms which has the potential for wide range of applications in various domains such as urban microclimate monitoring. IoT-based sensor node will facilitate sensor data collection from large number of collaborating sensors installed at multiple locations of a city [11].

Postolache et al. implemented an air-quality measurement sensor network which had an array of air-quality sensors connected to the sensor node [49]. These sensor nodes were further connected to an acquisition center which in turn was connected to a central monitoring unit (CMU). The CMU and sensor node communicated using a router. This sensor network was based on intranet paradigm. The sensor nodes and CMU needs to be within the range of the router. In urban microclimate, the sensor nodes might be required to be installed at different locations with varying distances between them. Due to high-rise buildings and other physical barriers, such city-wide intranet sensor network implementation might not be feasible. Hence, cellular communication can be used to acquire the data from the sensor nodes.

Merlino et al. [41] observed that several research activities are carried out for infrastructure issues related to IoT and Cloud for Smart Cities but there is a lack of framework to support the services in the infrastructure provisioning. They proposed to extend the OpenStack framework to manage the sensing and actuation devices in order to fill the gap between Smart City application requirements and the underlying infrastructure. OpenStack is a framework for management of Cloud computing resources. Merlino et al. [41] implemented the Stack4Things solution in infrastructure-oriented approach. The Stack4Things uses OpenStack as the underlying technology to manage the IoT devices through Cloud infrastructure management. The Stack4Things approach can be used in managing the resource provisioning as well.

IoT is also defined as a transition from Internet of content to the Internet of real-world objects. These connected objects can communicate between themselves or any other Internet-enabled device [39]. With the development of cheap and power-efficient chipsets in semiconductor industry, the Machine-to-Machine (M2M) communication is directing toward a future where billions of devices will be connected through a range of communication networks and cloud services. With the increasing mobile Internet service providers, cellular network coverage, scalable connectivity, decrease in power consumption of chipsets, and cost-effective chipsets, the environment for "IoT-driven ecosystem" is being prepared [72]. IoT is one of the fastest growing industries in the world [26]. The development of integrated modules of computing, sensing storage and communication in a low-footprint integrated software, and energy efficient and smaller hardware have enabled the design of IoT-based sensor nodes [26].

Researchers have demonstrated the use of GPRS cellular modules in conjunction with generic microcontroller [74] for Internet connectivity.

Such IoT enabling hardware support has multiple advantages to a sensor node such as:

1. Small dimensions as compared to routers and conventional Internet enabling hardware.
2. Interfaced with microcontrollers.
3. Built-in support for protocols relevant for Internet such as Transmission Control Protocol (TCP), User Datagram Protocol (UDP), File Transfer Protocol (FTP), Hyper Text Transfer Protocol (HTTP), etc.
4. Real-time data transmission capability.
5. Low power consumption.
6. Capability to upgrade sensor software over Internet.

While the hardware development is enabling IoT-based sensor nodes, there are developments in low-footprint software to adjust accordingly. The data generated from sensor nodes have to be methodically stored which facilitates the accessibility and analysis by end users of the data [22]. With time, the amount of such data collected will be excessively large to be stored locally in the sensor node. Hence, cloud-based storage solutions are gaining popularity. It removes the requirement to add extra physical memory in the sensor node. There are various lightweight protocols and software being developed to target IoT-based device requirements, such as Extensible Messaging and Presence Protocol (XMPP) [54], Message Queuing Telemetry Transport (MQTT) [47], MQTT For Sensor Networks (MQTT-SN) [46], and Constrained Application Protocol (CoAP) [58]. These protocols are easy to implement, lightweight, and open-source [46, 47, 54, 58]. It offers the conventional server/client architecture such that the client (i.e., sensor node) has the requirement to publish the data (i.e., weather data) and server disseminates it to the other interested clients (i.e., users of weather data) [47]. If we use cellular Internet connectivity, the cost is subject to usage of amount of data. This approach requires persistent Internet connection and might result in high data budget. Alternatively, the data can be collected periodically using native protocols like FTP or secure file transfer protocol (SFTP), etc. Various lossless compression techniques such as Huffman coding, Prediction by partial matching (PPM), Lempel–Ziv compression, etc., can be utilized to decrease the data budget [43, 71].

3 Design of IoT-Based Architecture for Sensor Node

In Sect. 2, the requirement, feasibility, and rationale behind using IoT-based architecture for urban microclimate monitoring was explained. In this section, challenges and generic approaches in the development of IoT-based architectures for sensor nodes will be discussed.

Design of IoT-based architecture involves integrating technologies from various domains. In a broader sense, the design of IoT-based architecture can be divided into two domains, i.e., hardware development and software development. In the case of hardware development, selection of sensors, computing unit, and communication module is involved. For software development, new Software Development Life

Cycle (SDLC) needs to be applied considering the constraints related to IoT-based devices.

In general, the IoT-based sensor node design has many open technical issues to be addressed:

1. Computing power,
2. Power budgets,
3. Bandwidth,
4. Hardware and software security, and
5. Lack of widespread WSN data collection standards.

While designing an IoT-based architecture, several issues need to be addressed properly for proper functioning of the sensor node.

To be specific, monitoring microclimate has further multiple challenges associated with it. To enlist few,

1. Sensor selection:
   There are various types of sensors to measure a particular weather variable. For example, to measure temperature, there are DHT22, DHT11, DS18B20, etc., as shown in Figs. 4, 5, and 8. To select appropriate sensor to measure a particular weather variable is one of the important aspects in sensor node design. One should consider accuracy, communication interfaces, range, operating frequency, etc., while selecting a sensor.
2. Computing hardware:
   There are two types of computing units used in sensor node, i.e., Single-Board Computer (SBC) and microcontrollers as shown in Fig. 3. The SBCs have operating system such as Linux or Windows. The drawback of using SBCs is that their power requirement is substantially higher than microcontrollers.
3. Interfacing between sensor and computing unit:
   The sensors come with a particular interfacing technique, and not all computing

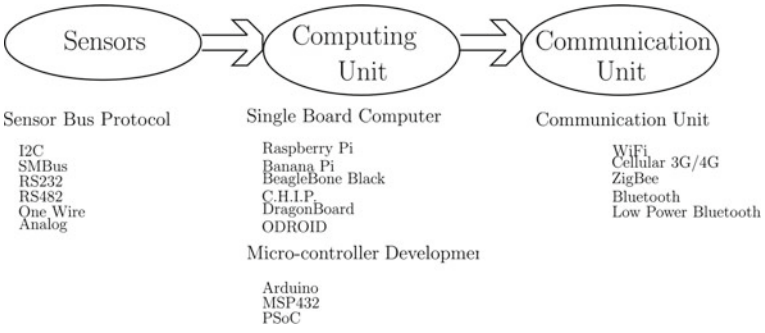


Fig. 3 Major components of a sensor node: sensor collects raw data and sends it to computing unit, which in turn processes it if required and then transmits to the base station using the communication unit

units support all the interfacing options. Hence, while selecting the computing unit and sensor, the compatibility of interfaces should be checked.
4. Internet of Things (IoT) enabling hardware: In urban setting, we can use 2G/3G cellular technique-based hardware such as SIM900, SIM5023, etc., to connect the sensor node to the Internet.
5. Power source: One of the important principles of IoT-based sensor node design is to operate with "minimum energy". This principle should be considered while creating IoT hardware and software components. If power requirement is low, various energy sources such as solar power, wind energy, etc., can be used in absence of power supply from the conventional energy grid to power the sensor node to operate independently.

In most cases, sensor nodes are battery powered. Hence to minimize the power dependency, the sensor node should be designed such that it consumes minimal power. In the following subsections, we will discuss in detail about the hardware, software, and IoT-based architectures.

3.1 Hardware

The typical hardware components of sensor nodes are sensors, computing unit, and communication unit, as shown in Fig. 3. During the design phase of a sensor node, every component should be selected such that it facilitates the final integration and avoids possible compatibility issues.

3.1.1 Sensors

To measure a particular weather variable, there are various sensors with different interfaces and varying accuracies. For example to measure ambient temperature, we can use DHT11, DHT22, DS18B20, STS30, etc.; to measure surface temperature, we can use MLX90621, MLX90615, etc., as shown in Figs. 4, 5, 8, 10, 6, and 7. These sensors differ in accuracy, interfacing technique, and power requirements as shown in Table 1 (Fig. 9).

One important aspect of any sensor integration in a sensor node is the hardware interface between the computing unit and the sensors. There are many hardware interfaces used by sensors such as,

1. Universal Asynchronous Receiver Transmitter (UART):
   UART is one of the oldest and most used protocols. Most SBCs and microcontrollers have embedded hardware UART [48]. It uses two separate data lines for transmitting and receiving data. It uses a simple protocol. The data packet is transferred between a low-level start bit and high-level stop bit. There is no fixed voltage level; one can use 3.3 V or 5 V. The speed of UARTs is relatively slow as compared to other interfaces.

Fig. 4 DHT11

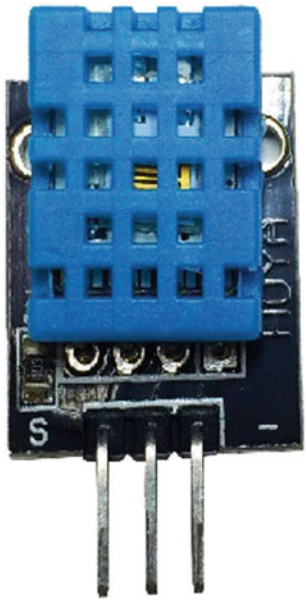


Fig. 5 DHT22

Fig. 6 MLX90621

Fig. 7 MLX90615

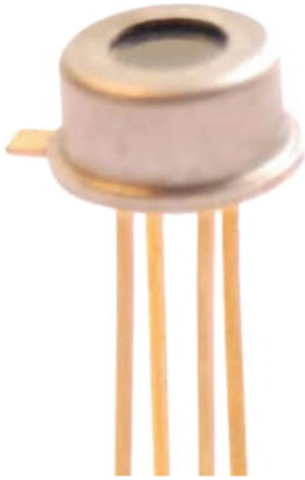


Fig. 8 DS18b201

Fig. 9 LM95231

Fig. 10 STS30

Table 1 Various temperature sensors

Sensor	Measures	Hardware interface	Accuracy
DHT11	Ambient temperature (AT) and relative humidity (RH)	Digital	±2 °C for AT and 5% for RH
DHT22	Ambient temperature (AT) and relative humidity (RH)	Digital	±0.5 °C for AT and 2–5% for RH
MLX90621	Surface temperature	I^2C	±1 °C
MLX90615	Surface temperature	I^2C	±1 °C
DS18B20	Ambient temperature	1-wire	±0.5 °C
LM95231	Ambient temperature	SMBus	±0.75 °C
STS30	Ambient temperature	I^2C	±0.3 °C

Temperature sensors like TMP104, TMP107; pressure sensors like TPIC83000-Q1; and obstacle-detection sensors like ultrasonic EV3 use UART interface.

2. Inter-Integrated Circuit (I^2C):
   I^2C is a synchronous interface protocol. I^2C is a two-wire protocol, i.e., one wire is for the clock and another is for the data [57]. It works on master–slave paradigm. The master and slave transmit and receive data over the same data wire, which is controlled by the master. The master creates the clock signal to synchronize the communication. I^2C does not use a Slave Select to select a particular device, but has 7-bit addressing space, i.e., one single I^2C bus can have 127 unique addressed devices.
3. System Management Bus (SMBus):
   SMBus is a two-wire bus. It is derived from I^2C for low bandwidth devices. Many sensors which have I^2C interface can be interfaced with SMBus of the microcontroller.
4. Serial Peripheral Interface (SPI):
   SPI is a simple serial protocol based on master–slave paradigm. The master sends a clock signal, and upon each clock pulse it shifts one bit out to the slave, and one bit in, coming from the slave. The master uses Slave Select (SS) signals to control multiple slaves on the bus [32].
   Sensors such as MS5803-14BA, LIS3DH, MPL115A1, LIS331HH, and MPU-6000 use SPI interface.
5. RS232:
   For long-distance communication up to 50 feet, the 3.5 V or 5 Volt UART is not reliable. Hence, the UART is converted to a higher voltage, typically +12 V for "0" and −12 V for "1". This modified UART is called RS232. There are RS423, RS422, and RS485, which differ in number of receiver/transmitter channels, maximum cable length, data rate, driver voltage, etc.
   Sensor such as HRLV-MaxSonar-EZ0, LV-MaxSonar-EZ0, etc., use RS232 interface.
6. 1-Wire: 1-Wire, as the name suggests, uses a single wire to transmit data. It is used for low-speed and low-power communication sensors.
   Sensors like DS18B20, DS18S20, MAX31820, and DS2760 use 1-wire protocol.
7. Analog:
   Many sensors reflect on the status of particular sensing variable based on the output voltage or current. These analog output voltage/current values are converted to digital readings by the computing unit based on the respective conversion equations normally obtained using a calibration process.
   Sensors like LMT84, DRV5053-Q1, IVC102, and OPT101 have analog interface.

3.1.2 Computing Unit

There are various available cheaper computing/controller units for development of IoT applications ranging from low-power 8-bit microcontrollers to 64-bit multi-core systems as the RaspberryPi 3. There are 8-bit microcontrollers (MCUs) such as Arduino, Teensy, etc., and 32-bit processors-based SBCs such as Raspberry Pi, Orange Pi, BeagleBoard Black, etc. These computing units are based on various architectures, such as MSP430, ARM, AVR, Cortex-M0, Cortex-M3, etc. Some examples of microcontrollers and Single-Board Computers (SBCs) are shown in Fig. 3.

1. Single-Board Computers (SBCs): SBC is a single Printed Circuit Board (PCB) board which has embedded microprocessor, memory unit, input/output channels, and other required features to work as a functional computer. The advantages of using an SBC are as follows:

   1.1 It has an Operating System (OS) running on it.
   1.2 It can handle multitasking operation.
   1.3 It supports high-level language such as C, Java, Python, and Perl. Therefore, software development time can potentially be decreased compared to conventional and platform-specific low-level programming languages used in embedded systems.

   SBCs have been gaining popularity due to their significant computing power and networking capabilities compared to traditional microcontrollers while maintaining non-prohibitive costs. There are many commercially available SBCs such as Raspberry Pi, C.H.I.P., and Beaglebone Black as shown in Figs. 11, 16 and 14.

   There are a lot of examples in literature showcasing the use of SBCs in WSNs. [3] presented an SBC-based sensor network system which had two parts, i.e., Sensor Node (SN) and Web Server (WS). In the system proposed by [3], SN functions as Data Acquisition Unit (DAU), consisting of the processor board while WS functioned as a storage unit. The SN and WS communicated using wireless network technology. In [38], Katsuyoshi et al. proposed a home energy management system based on an SBC. They demonstrated the applicability of SBCs in sensor data gathering. Radiation hardened SBCs are used in spaceflights, CubeSats, SmallSats, and satellite systems. In [40], a space-grade radiation hardened SBC at a substantially lower cost, lower power, and smaller form factor as compared to space-grade solutions available from aerospace manufacturers was presented.

   On the other hand, there are certain drawbacks of using SBCs such as:

   1.1 High power consumption compared to microcontrollers.
   1.2 High-level operating system cannot guarantee real-time operation.
   1.3 The majority of SBCs are mostly oriented toward computing power and network interfaces, resulting in a limited amount of available controlling peripherals found in microcontrollers (such as Analog-to-Digital Converters, Quadrature Encoders, etc.).

2. Microcontrollers:
   A microcontroller is defined as a single chip microprocessor which incorporates data memory, program memory, and input/output ports on a chip. A microcontroller is a compressed microcomputer designed to control the functions of embedded systems such as robots, home appliances, wearable gadgets, etc. This chip has integrated circuitry required for the Central Processing Unit (CPU), Arithmetic Logic Unit (ALU), and memory access arrangements components [59].
   There are multiple vendors for microcontrollers, such as Microchip, Cypress, Maxim Integrated, Intel, NXP, etc. There are various types of microcontrollers available in market with different word lengths such as 4 bit, 8 bit, 32 bit, and 64 bit such as AVR8, AVR32, MARC4, PSoC1, MPC500, HT32FXX, MCS-48, PIC, LPC800, MSP432, TMS370, and Stellaris.
   To use a microcontroller, we can either design our own customized board, or buy some development boards such as Arduino, MSP-EXP430G2, PSoC, etc. A microprocessor development board is a printed circuit board which has integrated microprocessor and some required support peripherals. Arduino is one of the popular development boards, and it contains of an 8-bit AVR microcontroller such as ATmega8, ATmega168, ATmega328, ATmega1280, or ATmega2560.
   Typically, microcontrollers do not have an operating system. Microcontroller programming is done in low-level languages such as assembly language, Basic C, mikroC, etc. The program is loaded in the microcontroller memory and whenever it is powered on, it executes the loaded program. Recently, a number of real-time, low-footprint operating systems have been developed for microcontrollers such as Contiki [17], TinyOs [6], MagnetOs [9], and SensorWare [14].

   2.1 SensorWare provides a lightweight scripting language which enables programming sensors such that it utilizes computation, communication, and sensing resources of the sensor nodes efficiently. SensorWare can be installed in XScale-based prototype sensor node platform.
   2.2 Contiki is an open-source, multitasking, event-driven lightweight operating system which supports dynamic loading and replacement of individual programs and services. This feature makes it apt for resource-constrained computing units. It is designed for networked embedded devices. It can be installed on a number of microcontroller architectures, including the Texas Instruments MSP430, Atmel AVRs, and Zilog Z80 microcontrollers [17]. Contiki's communication stack has support for a broad range of communication hardware. It supports IPv6 networking and a typical installation requires less than 10 KB of RAM and 30 KB of ROM.
   2.3 TinyOS is an open-source, component-based, flexible, application-specific operating system for microcontrollers with limited resources. It supports event-centric concurrent applications and low-power operation design. It is written in Network Embedded Systems C (nesC), a variant of the C programming language. It supports a range of different microcontroller such as

AVR family of 8-bit microcontrollers, Texas Instruments MSP430 family of 16-bit microcontrollers, ARM cores, Intel XScale PXA family [6, 52], etc.

2.4 FreeRTOS is a real-time OS from Real Time Engineers Ltd., which can be installed on multiple microcontroller architectures such as ARM Cortex-M7, ARM Cortex-M3, AVR, PSoC, XMC1000 MSP430, etc. It supports multiple threads and software timers. It is written mostly in C programming language. It has a very small memory footprint of 6 K to 12K ROM, low overhead, and fast execution [10].

2.5 RIOT OS is designed to address the requirements of IoT devices. It allows standard C and C++ programming, provides multi-threading, real-time capabilities, and requires less than 1.5 KB RAM and less than 5 KB of ROM [8].

2.6 Brillo OS is an operating system released by Google under the brand of Android. This OS requires 32–64 MB of RAM to run and it can be used smart home appliances such as TVs, refrigerators, light bulbs, and sensors [70].

TinyOS is a better fit when resource preservation is a priority, whereas, Contiki is selected when higher flexibility on the system's control is preferred [52]. Meanwhile, SensorWare targets platforms with higher, but still limited, resources than the aforementioned OSs [17].

IoT targeted operating systems are being launched by an increasing number of mainstream companies such as Huawei (LiteOS) and Microsoft (Windows 10 IoT). Based on the available computing power, memory, and sensor types, one should select the appropriate operating system.
There are few drawbacks of using microcontrollers such as:

2.1 Low computing power as compared to SBCs.

2.2 Only support low-level programming languages such as C, Basic C, Assembly, etc., as compared to SBCs.

On the other hand, low-power microcontrollers are one of the most important enabling technologies for IoT-based sensor node development.

3. Comparison of various SBCs and microcontrollers:
There are a number of different parameters on which a comparison between SBCs and MCU development boards can be based.

Various factors have been selected as a basis of comparison and are presented in the following tables. Cost and weight data are presented in Table 2, power requirement and computing power in Table 3, input–output interface support in Tables 4 and 5, and programming support in Table 6 (Figs. 12, 13, 14 15, 16, 17, 18, 19, 20 and 21).

From Table 2, we can observe that there SBCs and microcontrollers have comparable dimension, weight, and cost.

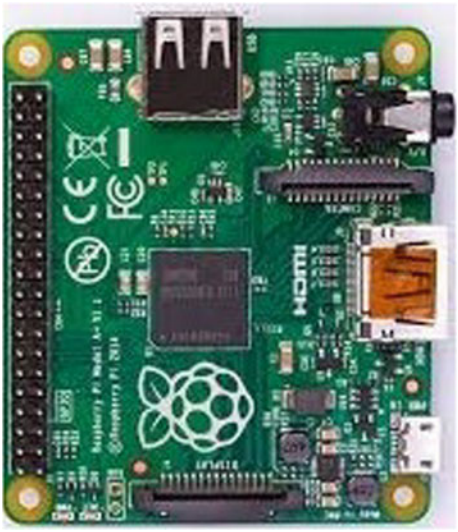


Fig. 11 Raspberry Pi A+

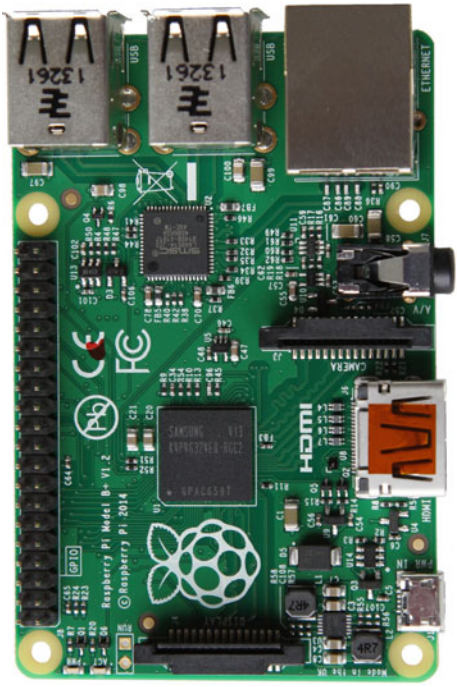


Fig. 12 Raspberry Pi B+

Fig. 13 Raspberry Pi zero

Table 2 Size, weight, and cost comparison

Name	Size (mm)	Weight (g)	Cost per unit US$
Single-board computer			
Raspberry Pi	85.6 × 53.98	45	25–35
BeagleBone black	86.3 × 53.3	39.68	45
C.H.I.P.	32 × 47.60	40	9–16
Microcontroller development board			
Arduino Uno	68.6 × 53.4	25	24
Arduino Mega	102 × 53.3	37	45
MSP-EXP430G2	75 × 59	20	10

Fig. 14 BeagleBone black

Table 3 Power requirement and computing power

Name	Processor	MIPS	RAM	Input voltage
Single-board computer				
Raspberry Pi	ARM BCM2835	1822–2451	256–512 MB	5 V
BeagleBone Black	AM335 × 1GHz ARM? Cortex-A8	2000	512 MB	5 V
C.H.I.P.	R8 ARM Cortex-A8	2000	512 MB	2.9–6 V
Microcontroller development board				
Arduino Uno	ATmega328P	1	32 KB	7–12 V
Arduino Mega	ATmega2560	16	256 KB	7–12 V
MSP-EXP430G2	MSP430	8	128–512 KB	5 V

Fig. 15 pcDuino

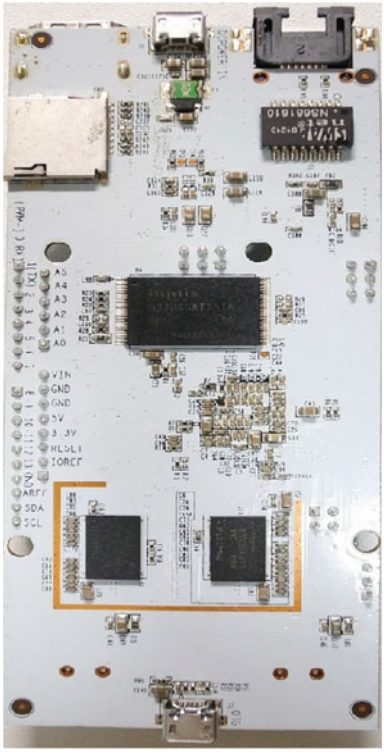


Table 4 Expansion connection

Name	Analog inputs	Digital I/O pins	USB ports
Single-board computer			
Raspberry Pi	0	14	1–4
BeagleBone Black	8	29	1
Microcontroller development board			
Arduino Uno	6	14	0
Arduino Mega	16	54	0
MSP-EXP430G2	8	8	0

From Table 3, we can observe that SBCs have more RAM as compared to the microcontrollers.

In Table 4, it can be observed that Raspberry Pi does not have any analog input pin. In such cases, one can use Analog-to-Digital Converter (ADC) to convert the analog input from analog sensors to digital output. Most of SBCs and microcontrollers have plenty of analog and digital pins. There are specific functions associated with some pins. For example, in Raspberry Pi B+ model, general-purpose input–output (GPIO)

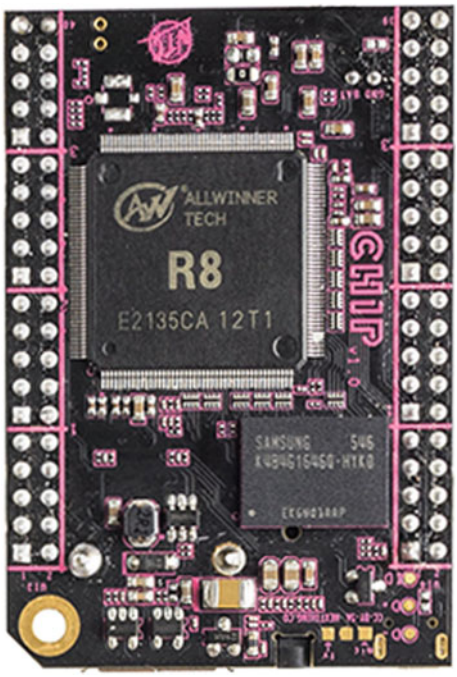


Fig. 16 C.H.I.P.

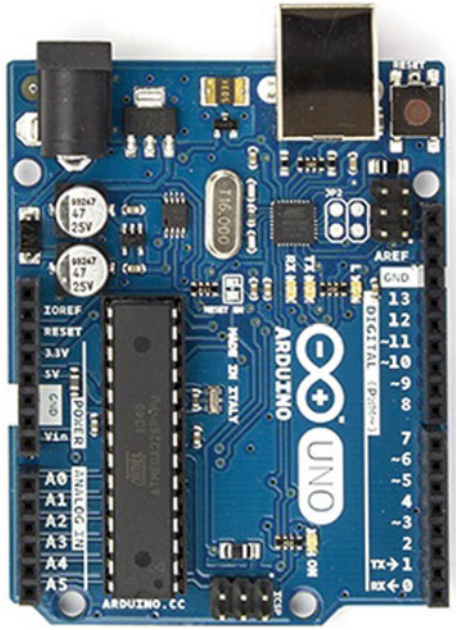


Fig. 17 Arduino Uno

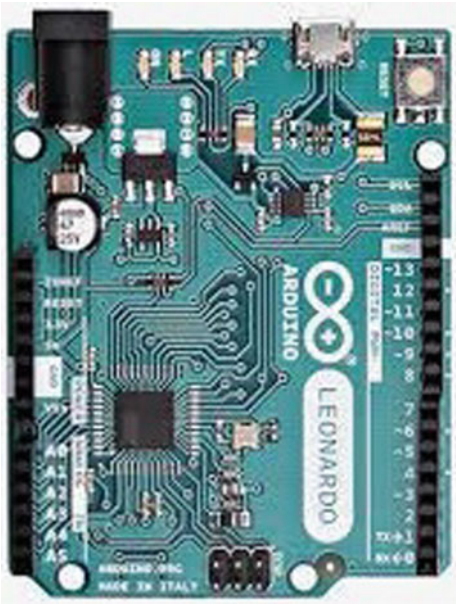


Fig. 18 Arduino Leonardo

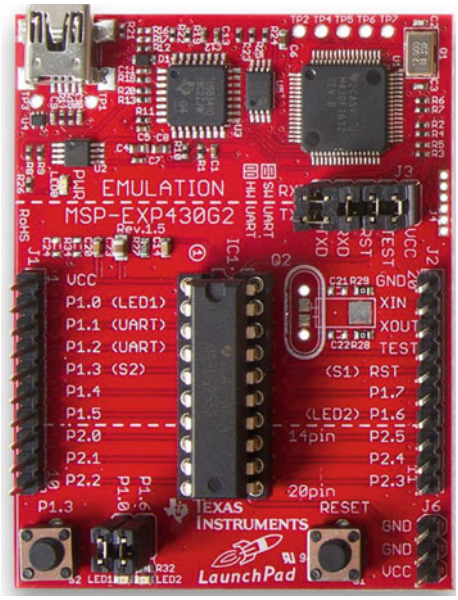


Fig. 19 MSP-EXP430G2

Fig. 20 Arduino Mega

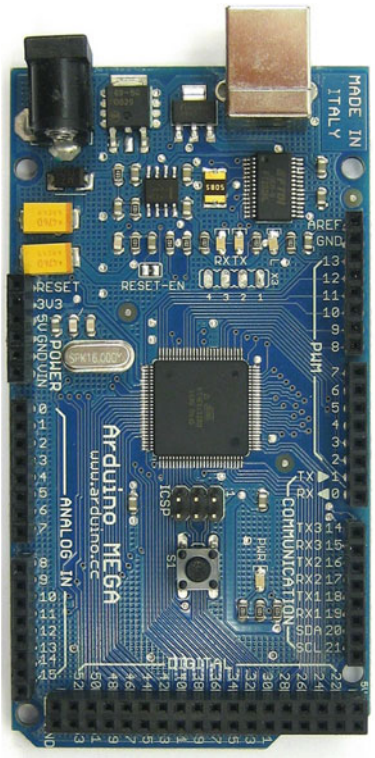


Table 5 Input–output interface support

Name	SPI	I^2C	RS232	UART
Single-board computer				
Raspberry Pi	2	2	0	2
BeagleBone Black	2	2	0	1
C.H.I.P.	2	0	0	2
Microcontroller development board				
Arduino Uno	1	1	0	1
Arduino Mega	1	1	0	4

pins 2 and 3 are assigned for I^2C interface; GPIO 9, 10 and 11 are assigned for SPI hardware interfaces.

In Table 5, it can be observed that the SBCs and microcontrollers support various hardware interfaces.

In Table 6, we can observe that the SBCs support different variants of Linux distribution such as Debian, Fedora, Ubuntu, etc. Some of the SBCs do support IoT

Fig. 21 PSoC®4

Table 6 Operating system and programming support

Name	Board operating system	Programming language/integrated development environment (IDE)
Single-board computer		
Raspberry Pi	Raspbian, Debian, NetBSD, Windows 10 IoT Core	C, C++, Java, Phyton
BeagleBone Black	Debian, Ubuntu, Fedora	C, C++, Java, Python, etc.
C.H.I.P.	Debian	C, C++, Java, Python, etc.
Microcontroller development board		
Arduino Uno	TinyOS, Contiki	C++, C, JArduino, Arduino
Arduino Mega	TinyOS, Contiki	C++, C, JArduino, Arduino
MSP-EXP430G2	TinyOS, Contiki	C, Energia

variation of Windows. The microcontroller boards can be programmed using the IDEs like Arduino, Energia or programming languages such as JArduino, C or C++. JArduino is customized java implementation for Arduino microcontrollers. Arduino development boards are based on Amtel AVR microcontroller which is supported by TinyOS and Contiki operating system.

3.1.3 Communication Module

The core requirements of IoT communication modules include low-power, IP-enabled, security, and reliability [20]. There are various standardization bodies working toward the creation of IoT-specific communication protocol stacks taking into account the constrained resources of IoT devices. There are many wireless protocols, such as ZigBee, IPv6 over Low-Power Wireless Personal Area Network (6LoWPAN), WirelessHART, IEEE 802.15.4e, ISA100.11a, etc. Among these, commonly used ones are Wi-Fi, ZigBee, and Bluetooth.

1. Wi-Fi:
   Wi-Fi is a wireless networking technology based on the IEEE 802.11 standard. It runs at 2.4 GHz range and supports speeds of 1 to 11 Mbps. There are various Wi-Fi modules available for microcontrollers and SBCs as shown in Figs. 22, 23, and 24.
2. Bluetooth:
   Bluetooth is a wireless protocol designed for short-range, low-power communication. It is mostly used to avoid cabling for communication for cell phone and computer peripherals such as headphone-audio, keyboard, mouse, and printers. It is based on IEEE 802.15.1 standard. It enables electronics devices to communicate with each other over a short range of up to 10 m at a data rate of 720 Kbps. It operates on 2.4 GHz unlicensed ISM (Industrial, Scientific, and Medical) band. There are various Bluetooth modules available for microcontrollers and SBCs as shown in Figs. 25 and 26.

   There are modules which have Wi-Fi and Bluetooth inbuilt as shown in Fig. 27.
3. ZigBee:
   ZigBee is a standard which defines Wireless Personal Area Network (WPAN) for a low data rate and short-range wireless networking. It is based on IEEE 802.15.4 standard. It has comparatively low power consumption as compared to other short-range networking technologies such as Wi-Fi and Bluetooth. The ZigBee module is shown in (Fig. 28).

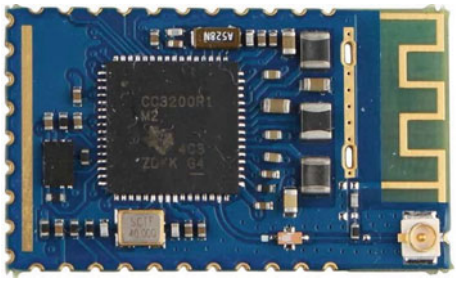


Fig. 22 Wi-Fi module

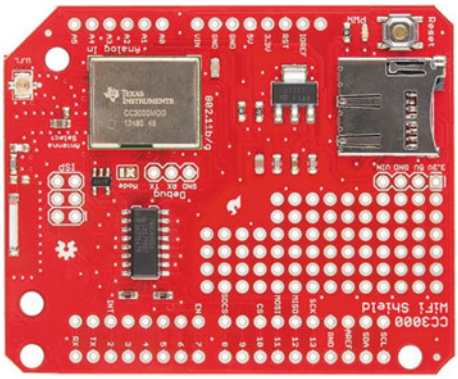


Fig. 23 Arduino Wi-Fi shield

Fig. 24 USB Wi-Fi dongle

Fig. 25 Arduino Bluetooth module

Fig. 26 Bluetooth USB

Fig. 27 Integrated Wi-Fi and Bluetooth module for Raspberry Pi

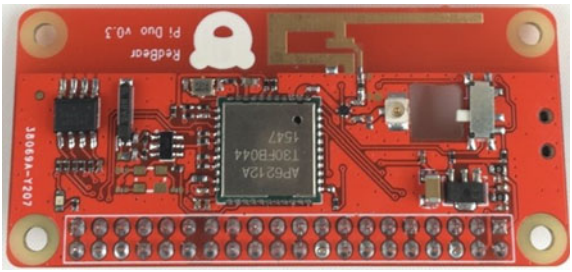
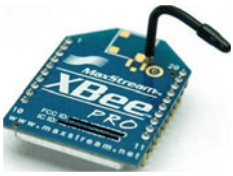


Fig. 28 ZigBee



3.2 Software

Software is a crucial aspect of the operation and behavior of a sensor node. There are various challenges, requirements, and development methodology associated with software for IoT devices. Unlike conventional software development methodology, IoT software development methodology needs to adapt to the dynamic and multiple requirements related to constrained resources.

3.2.1 Requirements and Challenges for IoT Software Development

From programming perspective, the sensor node software should acquire, process, and store the data or upload it to the cloud or server. As the IoT devices are resource constrained and are bound to encounter unexpected failures, the software should be resilient and able to handle unexpected hardware failures. The requirements and challenges of software requirement can be enlisted as follows:

1. Security and privacy: Data security and privacy mostly depend on software implementation and underlying protocols along with cryptography. Since IoT is among evolving technologies, best practices for some technologies are not yet

available. Also, IoT devices are resource constrained, so it is not recommended to use computing intensive security and privacy protocols. New protocols, easier to implement, are being proposed in the literature taking into account the resource-constrained nature of the IoT systems [16]. Furthermore, there is an increasing number of MCUs utilizing hardware accelerated encryption.

2. Middleware and Application Programming Interfaces (API) design: There are various IoT Middleware and APIs, which aim to integrate large numbers of heterogeneous real-world objects [67]. Middlewares and APIs are key design elements in IoT architectures.
   The middlewares and APIs, which offer a lot of features for interoperability, might be misused in unanticipated contexts; therefore, security concerns arise [69]. Hence, while creating such web services, APIs, and middleware, security poses as one of the core priorities.
3. Scalability: As the number of IoT nodes grow, the amount of data harnessed will also grow. This will add burden on the cloud/server and the underlying communication infrastructure. Hence, the large amount of data stream generated from IoT devices needs to be managed effectively and efficiently and software implementations should consider this aspect while building the software [67].
4. Resilient: The data collected from the sensor nodes might become invalid due to multiple reasons such as failure of sensor interface, drift in sensitivity and failure of sensors, etc. The IoT software should either try to notify such data corruption or discard the invalid data from the sensor data stream or try to apply corrective measures to the faulty data.
5. Power consumption: As the IoT devices are resource constraint, it is desired to design software based on "minimum energy" concept, i.e., the software should be designed such that it minimizes the overall consumption of the IoT system by efficiently controlling the switching between idle and operational states.

3.2.2 Software Development Life Cycle (SDLC)

An SDLC methodology is a collection of processes or methods which are applied for the development of a particular software depending on the project's aims, requirements, constraints, and goals. There are various SDLC models to address the various requirements and objectives of different types of project such as the waterfall methodology, spiral methodology, and iterative methodology.

1. Waterfall methodology: Waterfall methodology is one of the oldest methodologies in SDLC. In this methodology, the software is developed in a sequential, linear, and predictable process. It consists of six stages. Each stage is followed by the other as shown in Fig. 29. The stages are as follows:

   - Requirement analysis: In this stage, all possible requirements of the system to be developed are analyzed and documented in a requirement specification document.

- System design: Based on the requirement specification document from the previous stage, the system design is prepared. In this stage, the specification of hardware and overall system architecture is finalized.
- Implementation: Based on the requirement and system design, the various independent components of the system are developed in small units.
- Integration and testing: All the independently developed units are integrated and the overall functionality is tested.
- Deployment: Once the integration and the functional and nonfunctional testing are successfully completed, the product is deployed.
- Maintenance: Despite rigorous testing, there are some issues which arise when users start using the software. To fix the reported issues software patches or hot-fixes need to be released. Maintenance is done to deliver the upgrades or fixes to the users of the software.

All these stages are cascaded one after another such that if a phase is completed, it is not feasible to revisit it. This means that once the initial requirement is set, the new requirements would not be supported. Although this methodology provides stability [19], it does not offer the freedom to go back and make changes in the previous stages while developing the system [51]. In case of IoT systems, the technology is changing fast and needs to adapt to the continuously changing hardware, communication protocols, etc. As the waterfall methodology has been traditionally used for software development in mainstream computing systems, it seems to lack the required flexibility for the dynamic environment of IoT systems.

2. Iterative methodology: In the Iterative methodology, development of software begins by specifying and implementing major part of the software, which can then be reviewed in order to identify further requirements. Generally, if the overall major requirements of the complete system are clearly defined, this model is preferred. However, some functionalities or requested enhancements may evolve with time. This process is then iterated, producing a new version of the software for each cycle of the model.

   In this methodology, costly system *reconfiguration* or design issues may arise in later stages of development because not all requirements are gathered in the start of the system development. Due to the dynamic nature IoT systems, requirements of the complete system may not be clearly understood or defined. Hence, this methodology, in its original form, does not address all the required features for an IoT systems development life cycle.
3. Spiral methodology: The spiral model was proposed by Barry Boehm [12]. The development and testing go on an incremental basis as shown in Fig. 31.
   This methodology is useful for developing large and complicated projects. This methodology combines some aspects of iterative prototyping with the design sequences of the waterfall method. Based on the user evaluation, the development enters into next phase, and then again follows linear approach to implement

the requirements and feedback from the users.

There are four phases in the spiral methodology. These phases are as follows:

3.1 Identification: In this phase, the requirements are identified. As the development cycle moves on and comes back to this phase again, the software is reevaluated and new requirements are identified based on the interactions from the software users and input from previous phases.
3.2 Design: In this phase, the conceptual and architectural design based on the previous identification phase is created.
3.3 Construct or build: In this phase, based on design phase, the software development is done.
3.4 Evaluation and risk analysis: In this phase, the software developed in previous phase is evaluated and risk accessed.

4. Agile methodology:
In the 90s, a new SDLC methodology called "Agile Movement" [1] was published. Agile methodology is more flexible than traditional process methodologies. To develop reliable software for IoT systems at low cost, software development paradigms such as agile methodology [4] is recommended. Agile methodology works on the principle of incremental and iterative software development sequences as shown in Fig. 32.
Agile is composed of adaptive empirical small repeating cycles with short-term planning and constant feedback and inspection. In this methodology, the software developers can adapt to the changing requirements. This methodology facilitates collaboration with experts from various domains. There are various flavors of agile methodology such as Scrum [56], Kanban and Disciplined Agile Delivery [5].
5. Ignite and IoT methodology: For IoT-related software development, the methods for system engineering also need to be taken into account. In any IoT system, the software needs to be fine tuned and intelligently integrated with the hardware. There are some IoT-specific software development strategies discussed in Ignite—IoTMethodology [25, 61]. Based on the IoT project of various industries such as automotive, energy, and manufacturing, Slama et al. [61] proposed the'Ignite—IoT Methodology' as an IoT product management methodology. The aim of this methodology is to continuously develop IoT best practices and make them available to public in the form of a framework.
The IoT Methodology has two parts as shown in Fig. 33:

5.1 Ignite—IoT strategy execution: In this stage, the organizations to devise their IoT strategy. It has more to do with overall strategy for the organization.
5.2 Ignite—IoT solution delivery: In this stage, the product and project managers plan, build, and run the IoT project. The software development strategy is a part of this stage.

6. IoT-A methodology:
In [61], the authors argued that IoT projects are multidisciplinary, and hence,

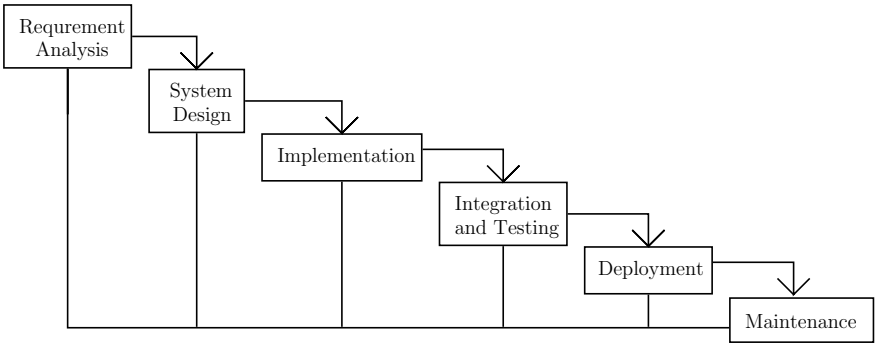


Fig. 29 Waterfall model

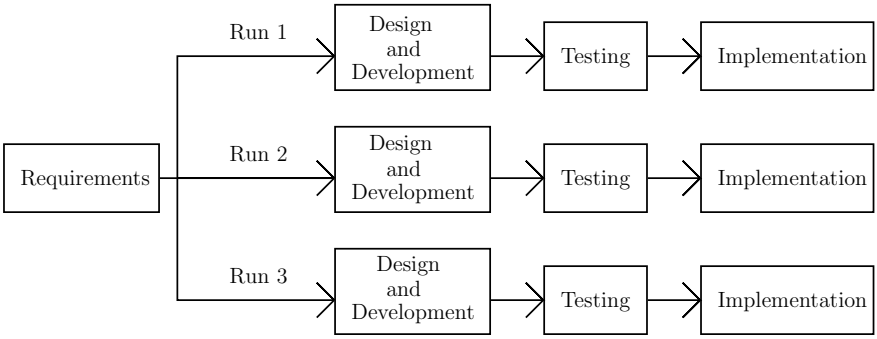


Fig. 30 Iterative methodology

their IoT methodology needs to combine multiple disciplines. Hence, the IoT methodology needs to be more interactive such that all the stakeholders are on the same page. Since IoT projects have to deal with an IoT of new technology, there is always a risk factor involved in this and to address this; a continuing review and analysis needs to be done. They proposed a "Plan-Build-Run" strategy, which can be integrated with various classic SDLC methodologies as shown in Fig. 34 Depending on the situation one of the approaches mentioned in Fig. 34 can be used but as a general rule, a generic Plan/Build/Run perspective should be applied to all of these different approaches for IoT-related project [61].

7. IoT-A methodology:
   In [51], the authors proposed a hybrid software development methodology combining relevant aspects from agile principles and the Spiral method. They argued that the rigidity of waterfall method renders it inappropriate for IoT-related software development. Since IoT is a fast-changing technology, the software might need to adjust while developing based on new requirements.
   The overall approach proposed by [51] can be enumerated as follows:

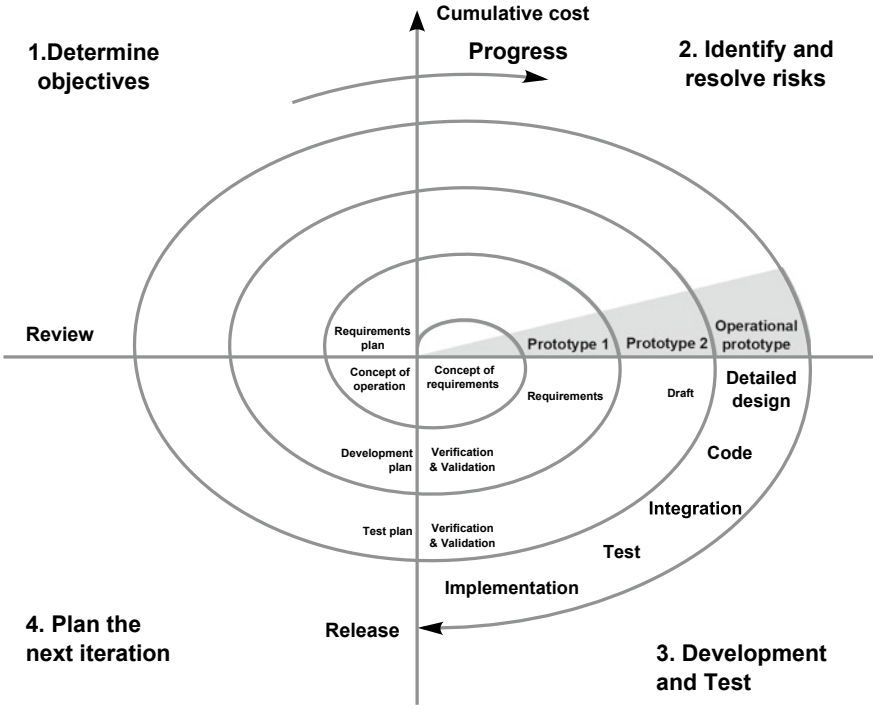


Fig. 31 Spiral model

7.1 The requirements are stored as a set of lists as shown in Fig. 35. As per the agile methodology, any new requirement or change request from the stakeholders can be added and implemented quickly.

7.2 To provide stability to the software development process, the requirements list remains unchanged for a certain time period.

7.3 At regular intervals, a review and final validation of the requirements list is done against the developed software with all the stakeholders. This is an aspect of Spiral software development methodology. Depending on the feedback of the stakeholder, change request, and priority of requirements, the further development is influenced and planned, thereby developing the software iteratively, as per the spiral model.

The process proposed in [51], all the primary features are developed first and tested. Based on first cycle of development, the second iteration is planned, thereby integrating the Spiral and Iterative methodology for IoT Software development.

There are many other IoT software development methodologies proposed by various researchers. In [73], Xie et al. proposed a new IoT SDLC approach which focused on reducing the design complexity and development cost of IoT application software.

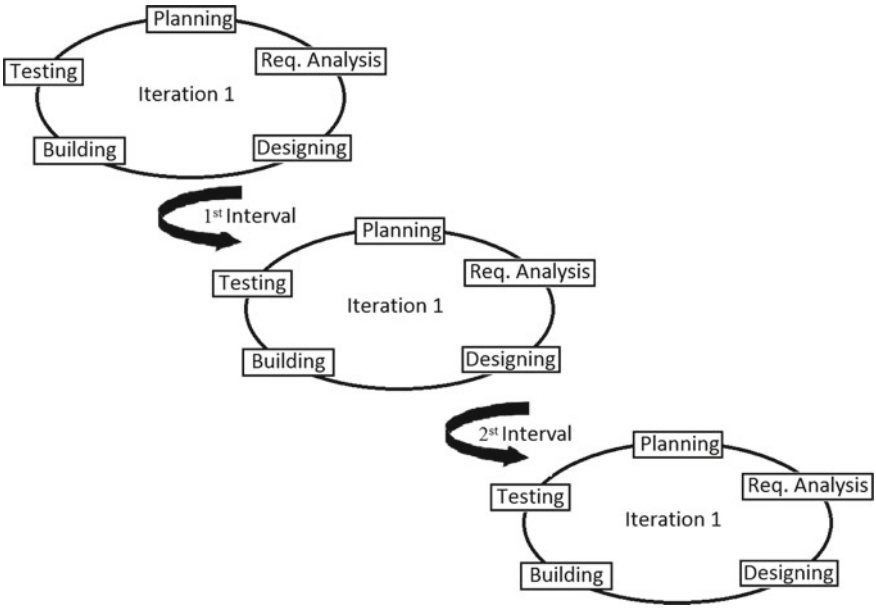


Fig. 32 Agile model

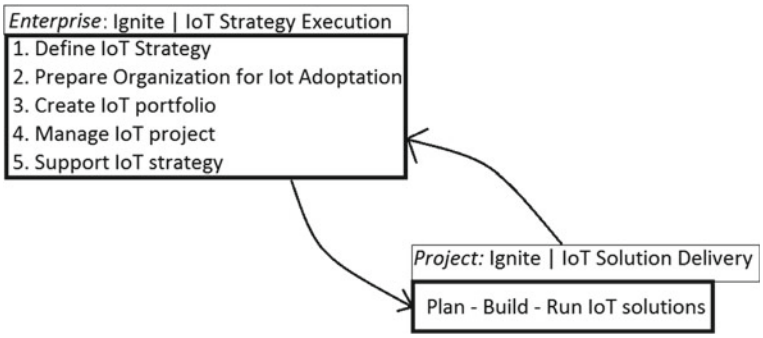


Fig. 33 Ignite | IoT methodology

Broadly categorizing, the IoT-based projects are integration of sensing/actuating components with computing and communication modules. The heterogeneity in the sensing/actuating, computing, and communication components leads to rise in complexity and thereby increasing the cost of developing of software. To address these various issues and develop the IoT-based software efficiently, we need to adapt to new emerging IoT centric SDLC methodology (Fig. 30).

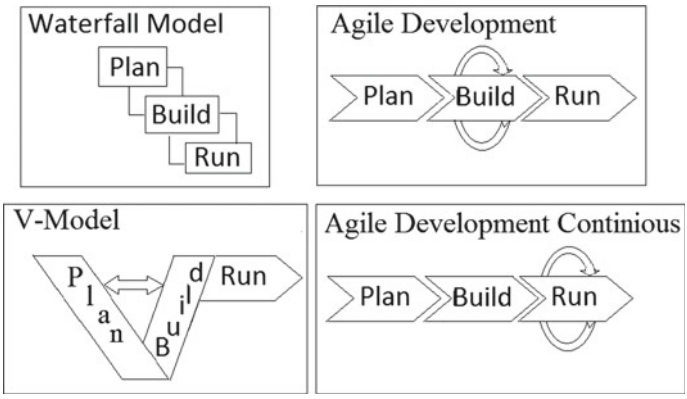
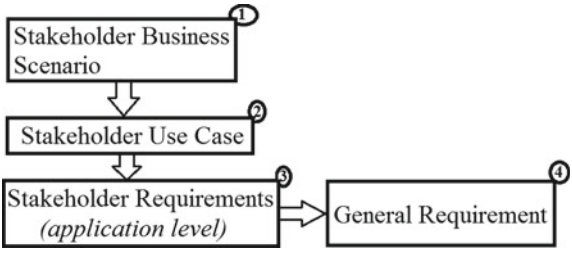


Fig. 34 Software development strategy proposed by ignite | IoT methodology

Fig. 35 General requirements process



3.2.3 Cloud-Based IoT Platforms

One approach to develop IoT-based sensor network is using the cloud-based centralized platforms such as Cumulocity [15], ThingWorx [18], and Xively. These global cloud-based platforms provide easy development, integration, and deployment of IoT applications. This approach spares the developer to maintain the server side cost. Some of the commercially available cloud-based IoT platforms are as follows:

1. Cumulocity: In this platform, sensor nodes act as agents and it connects to the cloud using RESfull HTTP APIs as shown in Fig. 36. Sensor nodes are treated as clients which can be accessed and manipulated. Users can be connected to the cloud and via cloud, the users can run commands on the devices.
2. ThingWorx: In this platform, targets application are integration through model-driven development as shown in Fig. 37.
   This platform supports CoAP, MQTT, REST/HTTP and Web Sockets [18]. The architecture of ThingWorx is shown in Fig. 37. It can be observed that ThingWorx is integrated with various cloud services such as Amazon Web Services (AWS), Salesforce, Twilio, etc., and also integrated to social services as Twitter, Facebook, Google Plus, etc.

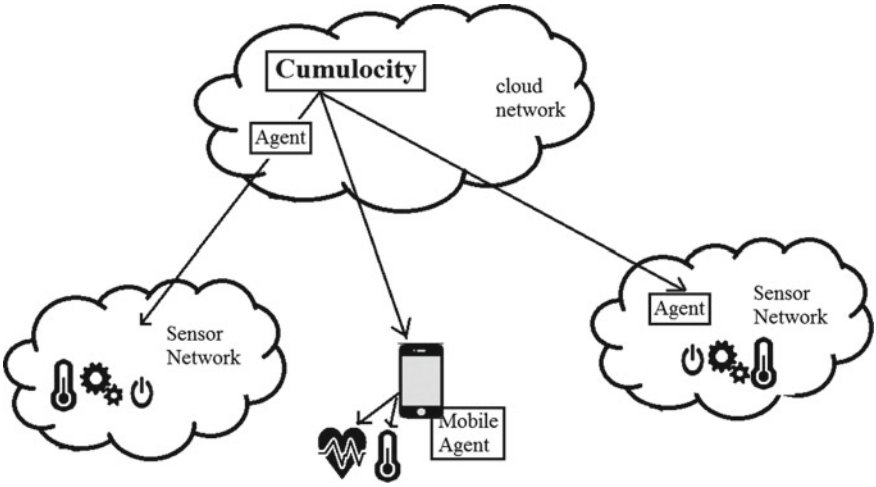


Fig. 36 Cumulocity architecture

3. Xively: This was formerly known as Pachube. It uses proprietary Xively API for MQTT, HTTP, and Web Sockets as a connection to the cloud. It has features to support fine grain access and works on client–server model as shown in Fig. 38. Each device is associated with unique ID and some form of authentication mechanism. Once the device is working, it can authenticate itself and upload time series data stream to the cloud and receive commands from the server. The users can also access the data via their smartphones, desktops, etc.
4. Watson IoT platform: It is an IBM IoT toolkit which integrates the device and management as shown in Fig. 39.
   This platform supports in collecting the data from the connected device and real-time analytics on the data collected. It provides communication to the IoT-device using the MQTT and TLS protocols.

3.3 Testbed for IoT Applications

A testbed is a platform for conducting rigorous and replicable testing of new technologies. IoT being one of the growing technologies which has potential for delivering multitudes of application in various domains; it is required to have testbeds to evaluate the applications developed for IoT. There are various issues that need to be tested

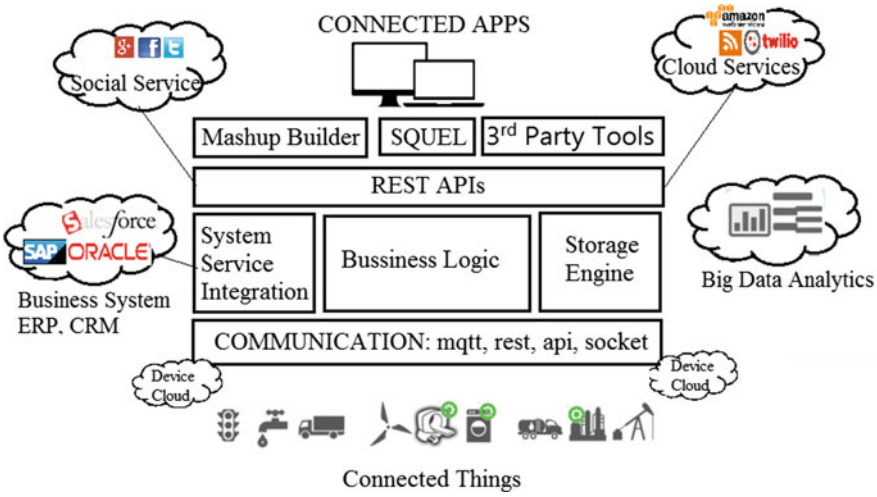


Fig. 37 ThingWorx platform overview

Fig. 38 Xively platform overview

before the IoT systems are deployed in large scale, such as reliability of wireless communications in dynamic environments, mechanism for allocation of constrained resources, maintenance of the system, etc. IoT solutions need to be thoroughly tested and fine-tuned before they are launched in the market. For robust development of IoT systems, a citywide testbed facility is required [23]. There are some discussions about how to construct robust and continuous IoT testbed platforms. Some researchers have proposed their own experimentation platform such as City of things [28], FIT IoT-LAB [2], and SmartSantander [55].

1. FIT IoT-LAB: It is an open-source testbed, which consists of 2728 low-power nodes and 117 mobile robots. These nodes and robots are available for experimenting of IoT technologies and applications, ranging from low-level protocols to advanced Internet services. This infrastructure accelerates the development

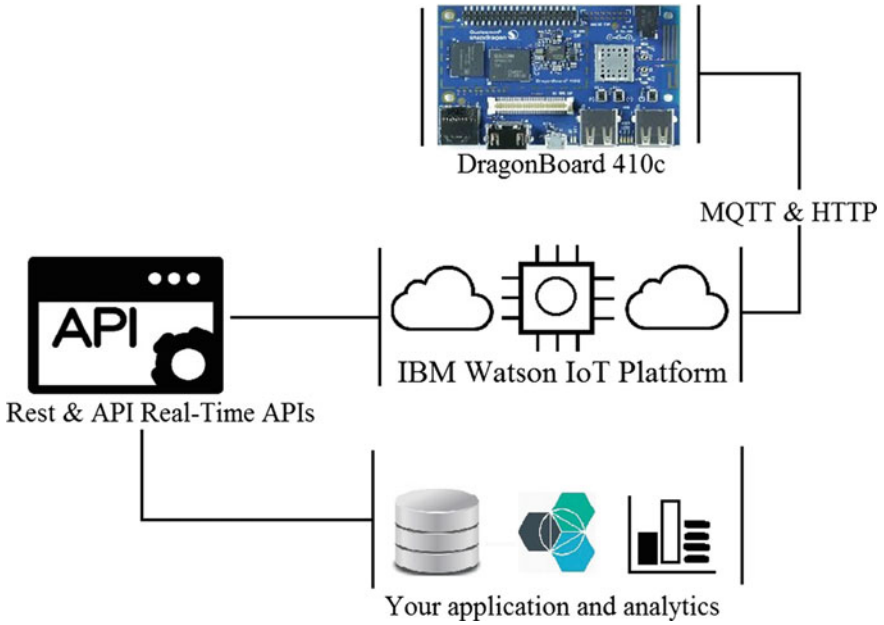


Fig. 39 Watson IoT platform overview

of IoT applications as it relieves the developer from the testing deployment of the application.

2. SmartSantander: It is one of the most advanced and active testbeds for IoT application testing. This testbed has around 20,000-node network. It offers a platform to conduct experiments at the scale of a city. The nodes are power constrained since batteries are used, therefore, they provide a real-life scenario of IoT nodes.
3. City of things: It is an integrated and multi-technology testbed for IoT experiments in urban areas [28]. It is integrated with JFed, i.e., a Java-based framework to support testbed; hence, it is compatible with a wide range of other testbeds such as SmartSantander. It can be effectively used to emulate large-scale multi-technology IoT networks and perform city-related big data experiments, etc.

These generic IoT testbeds can be further expanded with desired features for IoT-enabled urban microclimate applications, such as:

1. The hardware module should be able to withstand environmental harshness of the geographical region where it is to be deployed.
2. The testbed should provide mechanism to check the behavior of software under various critical conditions such as failure of communication module, failure of sensors, and its recovery.
3. The testbed should be able to measure the drift in the sensor reading over time.

4. The IoT testbed for urban microclimate should be able to provide the mechanism to measure the accuracy and precision of the sensors involved in the controlled environment.

These testbeds provide mechanisms to test the IoT-based urban microclimate monitoring software before actual deployment. Once the software is developed, it can be tested for scalability, resilience, and reliability. There could be multiple problems, bugs or unforeseen issues with software making testbeds an essential concept in developing IoT applications.

4 Implementation of Urban Microclimate Monitoring Using IoT-Based Architecture

Based on the IoT paradigm explained in the previous sections, a sensor network has been developed and installed in the city of Abu Dhabi, UAE [27]. The schematic diagram of the main portion of the typical sensor node is shown in Fig. 40, and a typical installation is shown in Fig. 41.

The design and development of IoT-based sensor node involve selection of sensors, integration of hardware, and development of relevant software.

4.1 Sensor Selection

The design process of a sensor node starts with the selection of climatic parameters one wants to measure followed by the other considerations such as accuracy, precision, cost, interface, etc.

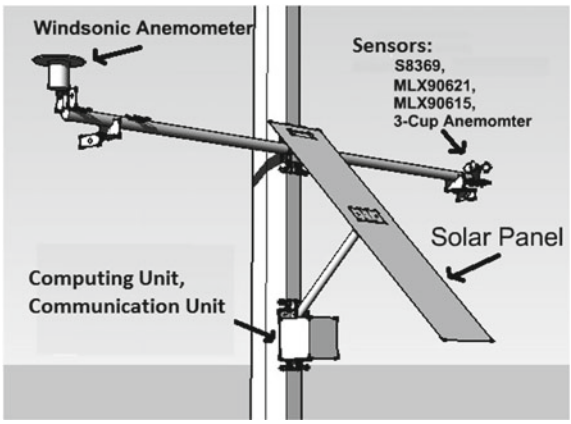


Fig. 40 Urban microclimate monitoring sensor node schematic

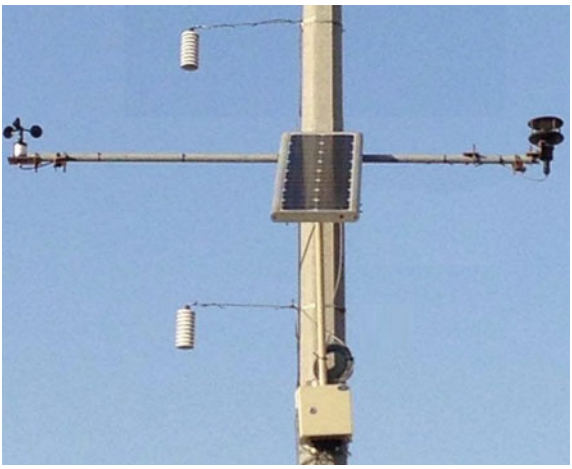


Fig. 41 Sensor node installed to monitor urban microclimate in Abu Dhabi

The modeling of the urban heat island relies on the application of the energy and momentum conservation principles to the urban canopy layer—the air volume immediately above the urban surfaces and extending approximately up to the height of the tallest buildings [45]. The microclimate model requires the various heat and momentum flux rates at the boundaries of the control volume (urban canopy layer). To determine the vertical fluxes of sensible and latent energy from the ground, it is also necessary to measure surface temperature, ambient temperature, and humidity at different heights. Ideally, the soil temperature needs to be measured, but this cannot be done systematically, for practical reasons. In [27], a single pixel infrared temperature sensor (MLX90615) was employed to measure land surface temperature.

The temperature measured at various heights will give us a temperature gradient, which can be correlated to the shear turbulence versus convective turbulence and the stability/instability of the urban boundary layer atmosphere.

Based on these requirements, in [27], the following sensors are used:

1. MLX90621: This sensor is used to measures the vertical thermal gradient of the buildings in the urban canyon.
2. MLX90615: This 1-pixel infrared temperature sensor is used to measure land surface temperature.
3. DS18B20: Multiple DS18B20 sensors are used to measure ambient temperature at various heights. This will give us a temperature gradient, which in turn can be used to calculate vertical heat flux exchange.
4. S8369: This sensor measures the Global Horizontal Irradiance (GHI). The GHI data be used to correlate with wind and temperature to establish the degree at which it affects other environment variables.
5. Wind sensors: We use Windsonic WS1 sonic anemometer for accurate wind measurements. In some of the nodes, we also employed 3-cup anemometers to reduce the cost. This will result in less accurate wind measurements.

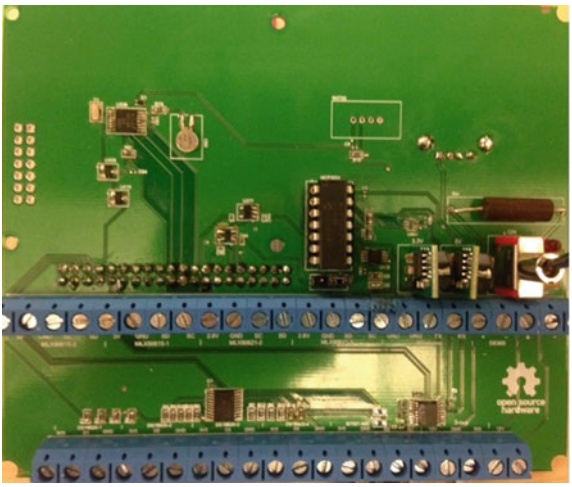


Fig. 42 Intermediate hardware between sensors and controller unit

4.2 Hardware Development

The hardware of any IoT-based sensor node has typically four components, i.e., computing unit, sensors, communication unit, and power source. There are various types of microcontrollers and SBCs, which can be used as control units. We need to select an appropriate control unit based on their ability to interface with selected sensors, power consumption, computing power, etc.

The senors with various types of hardware interfaces need to be integrated with the selected control unit. In [27], the IR sensors (MLX90615 and MLX90621) have I2C interface, DS18B21 has 1-wire interface, windsonic anemometer has RS232 interface, and S8369 outputs analog signal. To integrate various types of sensors with the computing unit efficiently, a hardware layer was introduced between the sensors and the computing unit as shown in Fig. 42.

4.3 Software Development

The software defines the behavior and operation of the sensor node. The IoT-related software revolves around data acquisition, data processing, and transferring the data to remote host over Internet. Generally, an IoT application consists of multiple sets of sensing and actuating components. The heterogeneity in hardware induces complexity in the development of software. The IoT software development differs from the conventional software development in many ways due to heterogeneity of hardware, reliability of connectivity to Internet, and multiple points of failure [66]. The multiple points of failure impose the requirement of handling and recovering from such failures on the software.

The desirable features of software of an IoT-based sensor node are as follows:

1. Acquire and process data from the sensors: The sensors and the computing unit can communicate over various hardware interfaces such as I2C, 1-Wire, digital or analog, etc. Some sensors directly give the actual value of the physical parameter it measures while some give a raw data which needs further calculation to get the actual corresponding value of the physical parameter. Based on its application, the computing unit can acquire the raw data from the sensor and process it, i.e., discard faulty reading and if required, calculate the actual corresponding value from the raw data.
2. Store data efficiently: The sensor nodes are generally resource-constrained devices, and they have a limited memory. The sensor nodes constantly generate high volume heterogeneous data. The data needs to be stored such that it meets the memory constraint of the sensor node.
   It is a good practice to periodically compress the data files, automate the deletion of unwanted log files, and generate a notification before a specified memory limit is reached. Also, if the data files are uploaded to the remote host, they can be safely deleted to prevent the sensor node from reaching the memory limit.
3. Upload data to server/cloud: The IoT-enabled sensor nodes are connected to remote host over Internet. The sensor data needs to be uploaded to the server periodically. There are various protocols such as FTP, SFTP, SCP, FTPS, or WebDAV to transfer files to remote host. Based on one's requirement and system capacity, the sensor data collected needs to be periodically uploaded to the server/cloud.
   The sensor data can also be streamed in real time to remote server/cloud but it will come at the expense of high bandwidth, power requirement, and increase in cost. Based on the application of the sensor node data, the data can be either streamed or periodically uploaded.
4. Recover from the sensor node failures: The software needs to handle the sensor node failure. The most important aspect of recover is to preserve the data and notify the cause of the failure.
5. Apply correction to the sensor data, if required: The sensors might show some drift after certain time of operation. In such scenario, it might not be possible to replace the sensor altogether. The software can handle this and apply correction to the data in reference to some standard reading.
6. Monitor sensor node health: The software should monitor various important aspects of the sensor node such as CPU usages, memory usages, power status, etc., to keep a track of the health of the sensor node. As the sensor nodes are bound to encounter unexpected failures, the software needs to monitor the health of the sensor node to investigate the reasons for such failure after recovery.
7. Notify unrecoverable failure of sensor node: Due to certain unforeseen reasons, if the sensor node stops uploading or communicating with the remote server, the remote server should notify about such unrecoverable failures.

Apart from the abovementioned desired features for the software for sensor node, it needs to take into account the resource constraint nature of IoT-based sensor node. It is advisable to keep the design of the software simple and minimalist.

There are various software development methodologies such as waterfall methodology, iterative methodology, spiral methodology and agile methodology, and IoT-A methodology as explained in Sect. 3.2.2. For IoT-enabled sensor node, it is better to use agile or IoT-A methodology because these methodologies offer adaptive, interactive, and constant feedback and inspection mechanism for the software development. As these methodologies facilitate collaboration between the experts of various domains, it is well suited for IoT-enabled sensor node.

In [27], agile methodology was used to develop the software. The software developed in [27], was able to validate and apply corrections, if required, to the sensor data, periodically update the data to server, monitor sensor health, etc.

4.4 Communication Module

The communication module is one of the most important aspects of sensor node. If implemented properly, the communication module can facilitate many desirable features such as:

1. No need to physically access the sensor node for data acquisition.
2. Using methods like reverse ssh tunnel one can connect to the sensor node remotely.
3. Periodically update and upgrade the software of the sensor node from Internet code repository like github, bitbucket, etc.

In cities, Wi-Fi and 2G/3G/4G cellular networks are readily available, and hence, Wi-Fi or cellular network can be used to connect the sensor node to Internet. For remote locations, where Wi-Fi or cellular network are not readily available, multiple sensor nodes can be connected to an Internet gateway using communication modules such as Bluetooth, Zigbee, etc. There are different types of communication modules explained in detail in Sect. 3.1.3.

It is imperative that the location of the sensor node should be such that it has a reliable coverage of cellular network if the sensor node is to use cellular network. In [27], 3G-cellular network was used to enable Internet connectivity in the sensor node. The 3G-cellular network was used to upload the sensor data to the server. In [27], Huawei E1750, a 3G-USB modem was used to provide Internet connectivity. In cities, it is advisable to use Wi-Fi or cellular network which can provide Internet connectivity.

One can also use Bluetooth, Zigbee, etc., as communication modules but these communication modules do not provide connectivity to the Internet. To overcome this, the sensor network designers can create a separate Internet gateway and connect multiple sensor nodes to the Internet gateway.

The operational duration of the communication unit might be detrimental for the power-constrained sensor node. To minimize the power cost, and data cost, the communication module can be kept operational duration as minimal as possible. Another important aspect of communication module is its failure handling. If the communication module of the sensor node fails, it will not be able to notify about the failure itself. In such scenario, the remote server should generate notification if a sensor is not communicating for a given threshold period of time. The software should monitor the behavior of communication module and report any discrepancy in its behavior.

4.5 Sensor Node Deployment

Once the hardware and software of the sensor node are developed, the sensor node needs to be installed in the desired locations. The deployment of sensor nodes involves the following factors:

1. Selection of locations: It is important to decide the locations for sensor node installation. The locations should be selected such that the measurements can represent the microclimatic profile. The locations in this study were selected based on a preliminary computational fluid dynamics model of the thermal flows in the city.

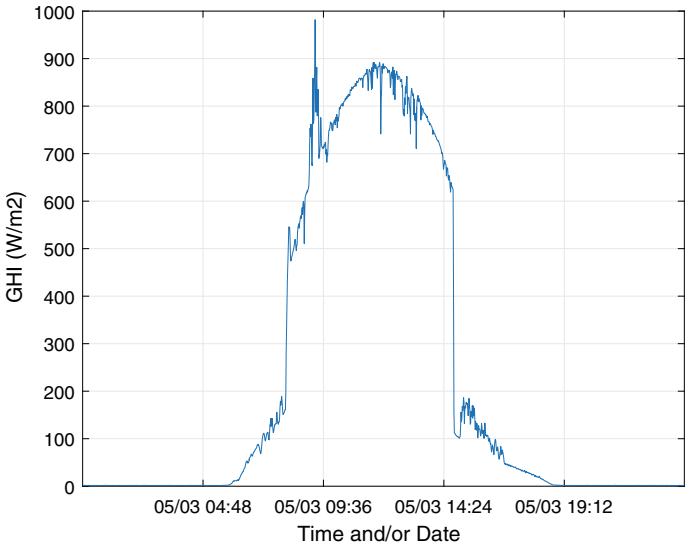


Fig. 43 Global horizontal irradiation

2. Permissions from the authorities: Since the urban sensor nodes are deployed in public areas, it requires getting permissions from the authorities based on the laws and policies of the land.
3. Logistic arrangement: Installation of sensor node requires needs proper organization, planning, and management. It requires trained manpower which can properly connect the sensor node.
4. Safety arrangements: The deployment of sensor node in urban setting might have some risk involved in it. The manpower involved in installation of sensor nodes should be provided with proper safety gears.

In [27], based on the schematic shown in Fig. 40, sensor nodes were designed and deployed at multiple locations of Abu Dhabi, UAE as shown in Fig. 41.

4.6 Data from a Sensor Node

The data from the sensor node is uploaded to server periodically. The data for a day from a particular sensor node is shown in the following graphs, Figs. 44, 43, 45 and 46.

The data collected over a period of time can help find trends, patterns, and correlation between various environmental variables of a microclimate. Figure 47 shows average daily temperature profile and average daily temperature gradient profile at three different heights from 07/09/2016 to 13/02/2017. Figure 48 shows the profile

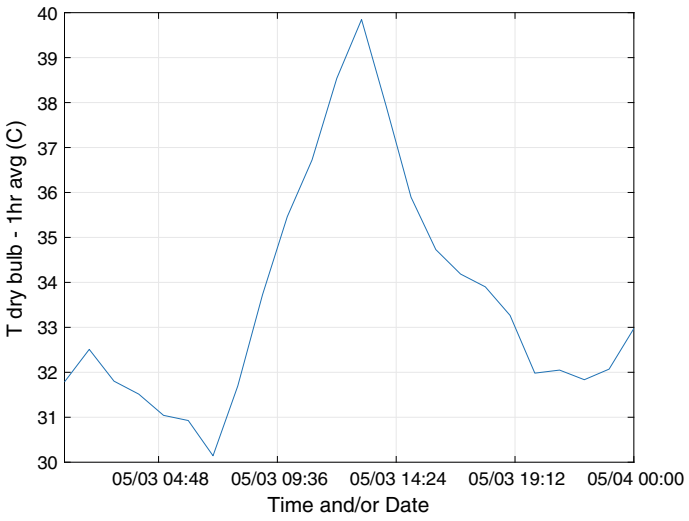


Fig. 44 Dry bulb temperature

of the average daily UHI intensity (urban minus rural) as well as day-averaged UHI intensity.

In connection with the Computational Fluid Dynamics (CFD) urban microclimate models, the measured data can be used for validation purposes of the models. Specifically, the simulated and measured values of the following quantities can be compared to evaluate the model:

1. Wind speed and direction,
2. Temperature at different heights, and
3. Urban surface temperatures.

The data collected from the various nodes will be used to validate the urban microclimate models. Also, the analysis of data can reveal insights which might be useful in microclimate weather forecasting. There are various such applications of this data.

5 Conclusion

In this chapter, we discussed the use of IoT-based architecture for urban microclimate modeling. A detailed overview of the architectures is provided and guidelines to implement the infrastructure to monitor the relevant environment variables are provided. The network developed in this work is now being used to develop advanced models for urban microclimate monitoring in Abu Dhabi city.

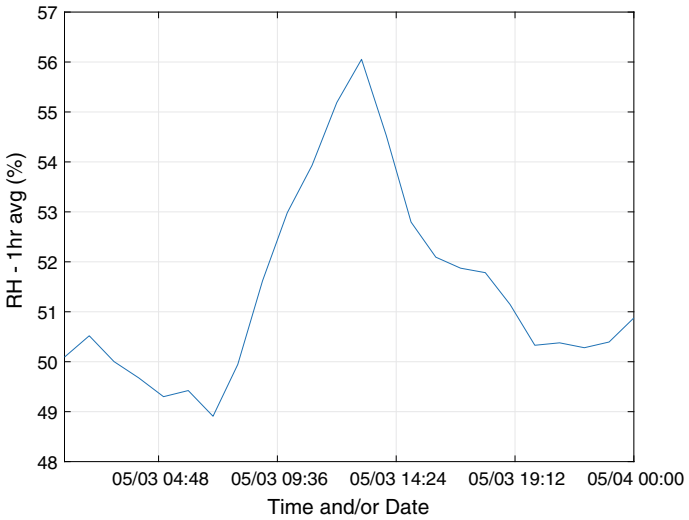


Fig. 45 Relative humidity

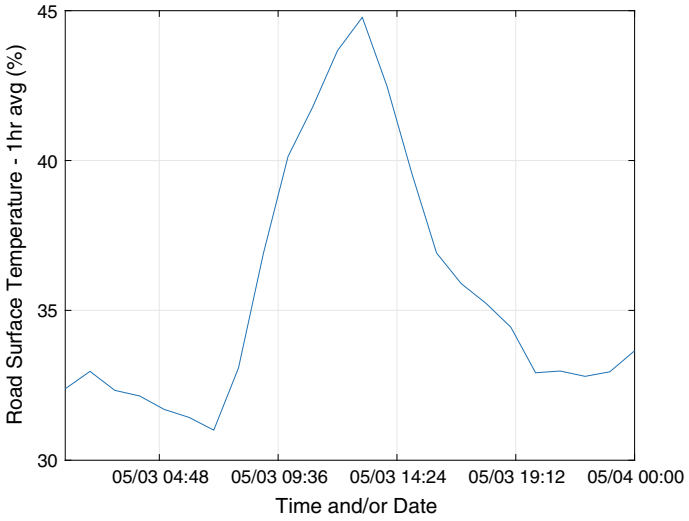


Fig. 46 Road surface temperature

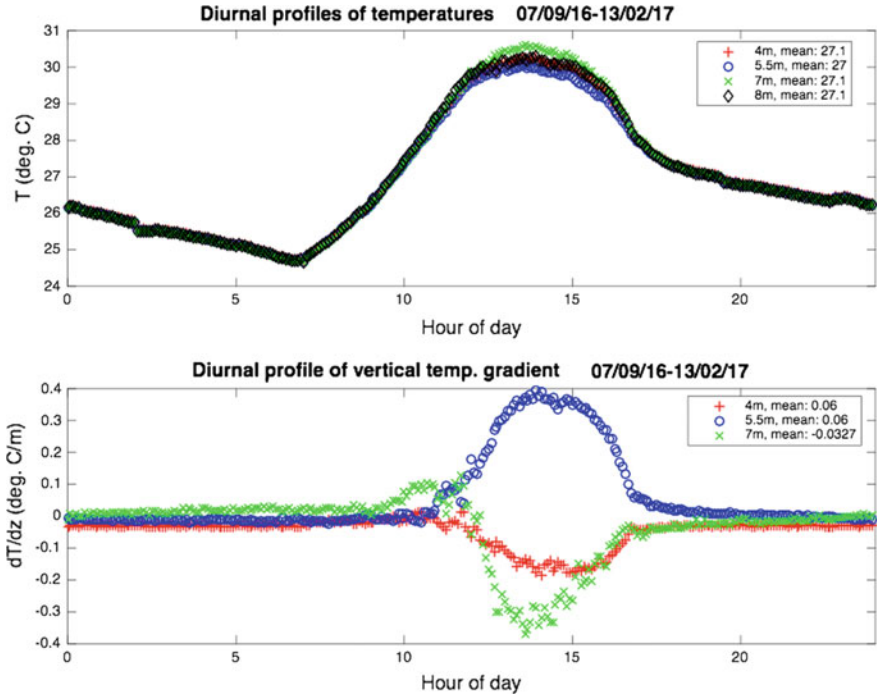


Fig. 47 Average daily temperature and gradient

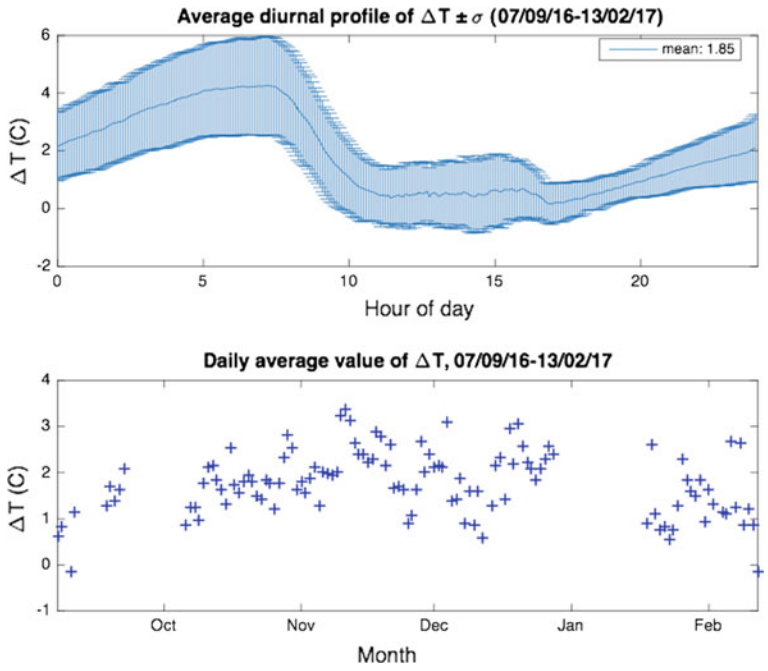


Fig. 48 Average daily UHI intensity (urban minus rural) and day-averaged UHI intensity

With growing IoT support framework, and advanced software development methodologies, we can custom design and develop IoT-enabled sensor nodes to effectively monitor urban microclimate. The advent of various standard sensors and their easy interfacing capabilities make it possible to develop custom sensor nodes instead of procuring expensive proprietary hardware and software for environment monitoring. Further, IoT paradigm connects the sensor node as a "Thing" to the Internet. This adds various features such as remote access to the sensor node via Internet, near real-time data acquisition, updating software over Internet, etc., to enable easy adaptation of hardware and software even after commissioning.

References

1. Abrahamsson, P., Warsta, J., Siponen, M.T., Ronkainen, J.: New directions on agile methods: a comparative analysis. In: 2003 Proceedings of the 25th International Conference on Software Engineering, pp. 244–254. IEEE (2003)
2. Adjih, C., Baccelli, E., Fleury, E., Harter, G., Mitton, N., Noel, T., Pissard-Gibollet, R., Saint-Marcel, F., Schreiner, G., Vandaele, J., et al.: Fit iot-lab: a large scale open experimental iot testbed. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 459–464. IEEE (2015)

3. Ahmad, R.B., Mamat, W.M.A., Mohamed Juhari, M.R., Daud, S., Arshad, N.W.: Web-based wireless data acquisition system using 32bit single board computer. In: 2008 ICCCE 2008. International Conference on Computer and Communication Engineering, pp. 777–782. IEEE (2008)
4. Alur, R., Berger, E., Drobnis, A.W., Fix, L., Fu, K., Hager, G.D., Lopresti, D., Nahrstedt, K., Mynatt, E., Patel, S., et al.: Systems computing challenges in the internet of things (2016). arXiv:1604.02980
5. Ambler, S.W., Lines, M.: Disciplined Agile Delivery: A Practitioner's Guide to Agile Software Delivery in the Enterprise. IBM Press (2012)
6. Amjad, M., Sharif, M., Afzal, M.K., Kim, S.W.: Tinyos-new trends, comparative views, and supported sensing applications: a review. IEEE Sens. J. **16**(9), 2865–2889 (2016)
7. Anagnostopoulos, T., Zaslavsy, A., Medvedev, A., Khoruzhnicov, S.: Top–k query based dynamic scheduling for iot-enabled smart city waste collection. In: 2015 16th IEEE International Conference on Mobile Data Management (MDM), vol. 2, pp. 50–55. IEEE (2015)
8. Baccelli, E., Hahm, O., Gunes, M., Wahlisch, M., Schmidt, T.C.: Riot os: towards an os for the internet of things. In: 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 79–80. IEEE (2013)
9. Barr, R., Bicket, J.C., Dantas, D.S., Du, B., Kim, T.W., Zhou, B., Sirer, E.: On the need for system-level support for ad hoc and sensor networks. ACM SIGOPS Oper. Syst. Rev. **36**(2), 1–5 (2002)
10. Barry, R.: FreeRTOS Reference Manual: API Functions and Configuration Options. Real Time Engineers Limited (2009)
11. Bellavista, P., Cardone, G., Corradi, A., Foschini, L.: Convergence of manet and wsn in iot urban scenarios. IEEE Sens. J. **13**(10), 3558–3567 (2013)
12. Boehm, B.W.: A spiral model of software development and enhancement. Computer **21**(5), 61–72 (1988)
13. Bolund, P., Hunhammar, S.: Ecosystem services in urban areas. Ecol. Econ. **29**(2), 293–301 (1999)
14. Boulis, A., Han, C.-C., Shea, R., Srivastava, M.B.: Sensorware: Programming sensor networks beyond code update and querying. Pervasive Mob. Comput. **3**(4), 386–412 (2007)
15. Cumulocity | connect to innovate. http://www.cumulocity.com/. Accessed 02 Apr 2017
16. Dalipi, F., Yayilgan, S.Y.: Security and privacy considerations for iot application on smart grids: Survey and research challenges. In: IEEE International Conference on Future Internet of Things and Cloud Workshops (FiCloudW), pp. 63–68. IEEE (2016)
17. Dunkels, A., Gronvall, B., Voigt, T.: Contiki-a lightweight and flexible operating system for tiny networked sensors. In: 2004 29th Annual IEEE International Conference on Local Computer Networks, pp. 455–462. IEEE (2004)
18. Enterprise iot solutions and platform technology. https://www.thingworx.com/. Accessed 02 Apr 2017
19. Friedman, A.L., Cornford, D.S.: Computer Systems Development: History Organization and Implementation. Wiley, New York, NY, USA (1989)
20. Gardašević, G., Veletić, M., Maletić, N., Vasiljević, D., Radusinović, I., Tomović, S., Radonjić, M.: The iot architectural framework, design issues and application domains. Wirel. Person. Commun. **92**(1), 127–148 (2017)
21. Georgakopoulos, D., Jayaraman, P.P.: Internet of things: from internet scale sensing to smart services. Computing **98**(10):1041–1058 (2016)
22. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of things (iot): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
23. Hamalainen, M., Tyrvainen, P.: A framework for iot service experiment platforms in smart-city environments. In: 2016 IEEE International Smart Cities Conference (ISC2), pp. 1–8. IEEE (2016)
24. G.K. Heilig. World urbanization prospects: the 2011 revision. United Nations, Department of Economic and Social Affairs (DESA), Population Division, Population Estimates and Projections Section, New York (2012)

25. Jacobson, I., Spence, I., Ng, P.W.: Is there a single method for the internet of things? (2016)
26. Jayakumar, H., Lee, K., Lee, W.S., Raha, A., Kim, Y., Raghunathan, V.: Powering the internet of things. In: Proceedings of the 2014 International Symposium on Low Power Electronics and Design, pp. 375–380. ACM (2014)
27. Jha, M., Marpu, P.R., Chau, C.-K., Armstrong, P.: Design of sensor network for urban microclimate monitoring. In: 2015 IEEE First International Smart Cities Conference (ISC2), pp. 1–4. IEEE (2015)
28. Latre, S., Leroux, P., Coenen, T., Braem, B., Ballon, P., Demeester, P.: City of things: an integrated and multi-technology testbed for iot smart city experiments. In: 2016 IEEE International Smart Cities Conference (ISC2), pp. 1–8. IEEE (2016)
29. Lazzarini, M., Marpu, P.R., Ghedira, H.: Land cover and land surface temperature interactions in desert areas: a case study of abu dhabi (uae). In: 2012 IEEE International Geoscience and Remote Sensing Symposium, pp. 6325–6328. IEEE (2012)
30. Lazzarini, M., Marpu, P.R., Ghedira, H.: Temperature-land cover interactions: the inversion of urban heat island phenomenon in desert city areas. Remote Sens. Environ. **130**, 136–152 (2013)
31. Lee, J., Baik, S., Lee, C.C.: Building an integrated service management platform for ubiquitous cities. Computer **44**(6), 56–63 (2011)
32. Leens, F.: An introduction to i^2c and *spi* protocols. IEEE Instrum. Meas. Mag. **12**(1), 8–13 (2009)
33. Li, T., Chen, L.: Internet of things: Principle, framework and application. In: Future Wireless Networks and Information Systems, pp. 477–482. Springer (2012)
34. Lund, D., MacGillivray, C., Turner, V., Morales, M.: Worldwide and regional internet of things (iot) 2014–2020 forecast: A virtuous circle of proven value and demand. International Data Corporation (IDC), Tech. Rep (2014)
35. Mainetti, L., Patrono, L., Stefanizzi, M.L., Vergallo, R.: A smart parking system based on iot protocols and emerging enabling technologies. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT), pp. 764–769. IEEE (2015)
36. Marceau, J.: Innovation in the city and innovative cities. Innovation Manag. Policy Pract. **10**(2–3), 136–145 (2008)
37. Miguel, M., Afshin, A., Armstrong, P.R., Norford, L.K.: A new validation protocol for an urban microclimate model based on temperature measurements in a central european city. Energy Build. **114**, 38–53 (2016)
38. Matsumoto, K., Yamagiwa, M., Uehara, M., Mori, H.: Proposal of sensor data gathering with single board computer. In: 2013 27th International Conference on Advanced Information Networking and Applications Workshops (WAINA), pp. 162–167. IEEE (2013)
39. Mattern, F., Floerkemeier, C.: From the internet of computers to the internet of things. In: From Active Data Management to Event-based Systems and More, pp. 242–259. Springer (2010)
40. Merl, R., Graham, P.: A low-cost, radiation-hardened single-board computer for command and data handling. In: 2016 IEEE Aerospace Conference, pp. 1–8. IEEE (2016)
41. Merlino, G., Bruneo, D., Distefano, S., Longo, F., Puliafito, A.: Stack4things: integrating iot with openstack in a smart city context. In: 2014 International Conference on Smart Computing Workshops (SMARTCOMP Workshops), pp. 21–28. IEEE (2014)
42. Naphade, M., Banavar, G., Harrison, C., Paraszczak, J., Morris, R.: Smarter cities and their innovation challenges. Computer **44**(6), 32–39 (2011)
43. Nelson, M., Gailly, J.-L.: The Data Compression Book, vol. 2. M&t Books. New York (1996)
44. Oke, T.R.: The energetic basis of the urban heat island. Q J. R. Meteorol. Soc. **108**(455), 1–24 (1982)
45. Oke, T.R.: Boundary Layer Climates. Routledge (2002)
46. OASIS Open: Mqtt for sensor networks (mqtt-sn), Oct 2016
47. OASIS Open: Mqtt version 3.1.1, Oct 2016
48. Osborne, A.: An Introduction to Microcomputers. Osborne & Associates (1978)
49. Postolache, O.A., Pereira, D.J.M., Girão, P.S.M.B.: Smart sensors network for air quality monitoring applications. IEEE Trans. Instrum. Meas. **58**(9), 3253–3262 (2009)

50. Price, J.C.: Land surface temperature measurements from the split window channels of the noaa 7 advanced very high resolution radiometer. J. Geophys. Res. Atmos. **89**(D5), 7231–7237 (1984)
51. Ref, D., Riedl, J., Heu, A.B.: Internet of things architecture iot-a project deliverable d6. 1-requirements list (2011)
52. Reusing, T.: Comparison of operating systems tinyos and contiki. Sens. Nodes-Operation, Netw. Appli.(SN), **7** (2012)
53. Rossi, C., Gaetani, M., Defina, A.: Aurora: an energy efficient public lighting iot system for smart cities. ACM SIGMETRICS Perform. Eval. Rev **44**(2), 76–81 (2016)
54. Saint-Andre, P.: Extensible messaging and presence protocol (xmpp): Core (2011)
55. Sanchez, L., Muñoz, L., Galache, J.A., Sotres, P., Santana, J.R., Gutierrez, V., Ramdhany, R., Gluhak, A., Krco, S., Theodoridis, E., et al.: Smartsantander: Iot experimentation over a smart city testbed. Comput. Netw. **61**, 217–238 (2014)
56. Schwaber, K.: Scrum development process. In Business Object Design and Implementation, pp. 117–134. Springer (1997)
57. Semiconductors, P.: The i2c-bus specification. Philips Semiconductors **9397**(750), 00954 (2000)
58. Shelby, Z., Hartke, K. Bormann, C.: The constrained application protocol (coap) (2014)
59. Sinclair, I.R.: Practical Electronics Handbook. Newnes (2000)
60. Singh, G., Singh, P.P., Singh Lubana, P.P., Singh, K.G.: Formulation and validation of a mathematical model of the microclimate of a greenhouse. Renew. Energy **31**(10), 1541–1560 (2006)
61. Slama, D., Bhatnagar, R.M., Morrish, J., Puhlmann, F.: Enterprise IoT. O'Reilly Media (2015)
62. Sobrino, J.A., Oltra-Carrió, R., Sòria, G., Bianchi, R., Paganini, M.: Impact of spatial resolution and satellite overpass time on evaluation of the surface urban heat island effects. Remote Sens. Environ. **117**, 50–56 (2012)
63. Sobrino, J.A., Jiménez-Muñoz, J.C., Paolini, L.: Land surface temperature retrieval from landsat tm 5. Remote Sens. Environ. **90**(4), 434–440 (2004)
64. Stewart, I.D.: Redefining the urban heat island. PhD thesis, University of British Columbia (2011)
65. Taha, H.: Urban climates and heat islands: albedo, evapotranspiration, and anthropogenic heat. Energy Build. **25**(2), 99–103 (1997)
66. Taivalsaari, A., Mikkonen, T.: A roadmap to the programmable world: software challenges in the iot era. IEEE Softw. **34**(1), 72–80 (2017)
67. Theodoridis, E., Mylonas, G., Chatzigiannakis, I.: Developing an iot smart city framework. In: 2013 fourth international conference on Information, intelligence, systems and applications (iisa), pp. 1–6. IEEE (2013)
68. Vakali, A., Anthopoulos, L., Krco, S.: Smart cities data streams integration: experimenting with internet of things and social data flows. In: Proceedings of the 4th International Conference on Web Intelligence, Mining and Semantics (WIMS14), pp. 60. ACM (2014)
69. Vukovic, M.: Internet programmable iot: On the role of apis in iot: the internet of things (ubiquity symposium). Ubiquity **2015**(November), 3 (2015)
70. Wee, A.: Google brillio os-another os dedicated to internet of things (2015)
71. Witten, I.H., Moffat, A., Bell, T.C.: Managing gigabytes: Compressing and indexing documents and images. IEEE Trans. Inf. Theory **41**(6), 2101 (1995)
72. Wu, G., Talwar, S., Johnsson, K., Himayat, N., Johnson, K.D.: M2m: From mobile to embedded internet. IEEE Commun. Mag. 49(4), 36–43 (2011)
73. Xie, K., Chen, H., Huang, X., Cui, L.: Low cost iot software development-ingredient transformation and interconnection. In: 2015 IEEE 21st International Conference on Parallel and Distributed Systems (ICPADS), pp. 44–51. IEEE (2015)
74. Yin, G., Jia, Z.-H., Wang, L.-J.: Study and design of wireless data communication experiment teaching system based on gprs. In: Education and Educational Technology, pp. 527–534. Springer (2011)

Models for Plug-and-Play IoT Architectures

Alexandros Tsoupos, Mukesh Jha and Prashanth Reddy Marpu

Abstract The Internet of Things (IoT) is already counting more than 15 billion devices connected to the web and over the following years, a rapidly increasing number of businesses and individuals are expected to become a part of this industry. In this context, enabling technologies and services are needed in order to accommodate this unprecedented interest. One of the major bottlenecks during the development of IoT products and/or services has been the vast diversity and incompatibility that exists among sensors, communication, and computation/controller modules. The various modules operating on numerous communication buses and protocols requires development of platform-dependent hardware and software drivers as well as vigorous testing from the developer side. This process is nontrivial and time-consuming and redirects the focus of developers from "what" to "how" to develop. In addition, users/developers are discouraged when they are obligated to perform tedious manual configurations before they are ready to use their products. Furthermore, there is a significant heterogeneity in IoT network architectures and existing automatic service discovery and configuration protocols. The majority of these protocols have been developed for conventional computer systems and as a result, it cannot be used by resource-constrained IoT devices. For the above reasons, various models of the well-known concept in mainstream systems of Plug-and-Play (PnP), are being introduced to the embedded systems world as well, to tackle the above issues. In the following chapter, an overview of what a Plug-and-Play architecture consists as well as a survey of the state of the art is presented.

A. Tsoupos · M. Jha · P. R. Marpu (✉)
Masdar Institute of Science and Technology, Masdar, UAE
e-mail: pmarpu@masdar.ac.ae

A. Tsoupos
e-mail: atsoupos@masdar.ac.ae

M. Jha
e-mail: mjha@masdar.ac.ae

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_5

1 Peripheral Plug-and-Play (PnP) Definition and First Attempts

The Plug-and-Play (PnP) concept, in terms of peripheral device integration, can be defined as the ability of the system to automatically detect and configure internal and external peripherals as well as most adapters. In personal computers, the manufacturers quickly recognized the tedious task of having to manually configure hardware jumpers and software settings, and started to shift towards architectures that would allow simple and automatic integration of peripherals.

1.1 NuBus

One of the first attempts of PnP architecture was the MIT NuBus that was introduced in 1984, and was first standardized in 1987 [1]. In comparison to the existing architectures at that time, featuring 8-bit or 16-bit buses, the Nubus was equipped with a 32-bit backplane to accommodate future systems. The NuBus did not feature a distinct bus controller which meant that all NuBus devices participated as peers to system control functions and arbitration. Furthermore, an identification scheme was present which allowed for automatic detection and configuration of NuBus cards from the host system. The bus offered a form of geographic addressing, meaning that each available slot had a small dedicated address with which was associated. Specifically, 32-bit physical addresses were multiplexed with the data lines and this address space was shared among all the existing slots. Up to 15 slots were available, and each of them featured a 4-bit ID field with which every communication process was qualified.

On the other hand, NuBus's implemented addressing scheme along and low clock speed (10 Mhz) compared to other architectures of the time, rendered the bus slow. Especially, the bus was not suitable for newer and faster I/O devices that did not have enough local buffering capabilities.

The NuBus is perceived as one of the pioneers of PnP hardware architectures. Texas Instruments acquired the project and developed a number of LISP and UNIX systems based on the NuBus. Later, after its standardization, the architecture was used in some Apple projects (Mac II, Mac Quadras) but eventually became practically extinct when Apple adopted the Peripheral Component Interconnect (PCI) bus [2] in their products.

1.2 MSX Bus

Another PnP architecture that was introduced in the 1980s, was Microsoft's MSX [3]. MSX was developed with the aspire to become the single industry standard for

home computing systems. The idea was to enable hardware peripherals, software applications, and computer systems, which is developed by different manufacturers/organization to be interoperable when they were MSX compatible.

The MSX system was based on the Zilog Z80-family of CPUs which is intended for home computing systems. MSX offered a very well-developed hardware abstraction layer which was implemented in the MSX-BIOS. This abstraction offered extensibility, peripheral independence, and instant PnP with zero user intervention. Virtual addressing was implemented using various slot/subslots which avoided any possible conflicts. The required drivers were already installed in the cards ROMs and were able to be automatically configured. The MSX was mostly popular in Japan and acted as the platform for many important Japanese game studios.

1.3 Micro Channel Bus

Last, the Micro Channel Architecture [4, 5] that was introduced by IBM in 1987, was the successor to IBM's ISA bus and proved to be the precursor to PnP systems known to current date such as the widely adopted PCI bus. The Micro Channel cards were 32-bit but also allowed 16-bit implementations for back compatibility and featured a unique 16-bit identifier which was software read. The OS/BIOS after reading the identifier successfully, proceeded with the search for appropriate device drivers. The IDs were stored in *Reference Disks* which IBM had to update in a regular basis. When a new card was inserted, and the system was unable to find corresponding drivers boot failures occurred. In the reference disks, along with the drivers further information was provided, such as the card's memory addressing and interrupts, which is crucial for the functionality of the system.

Using this information, the system could configure a new card without any intervention from the user. However, every time a new card was installed, the system changes (interrupts, etc.) had to be saved to a floppy disk which then became necessary for every subsequent hardware change. This proved to be an important design flaw, especially when the bus was used by large corporations. Therefore, although the Micro Channel was considered successful, soon after the release of the PCI bus, it became obsolete.

Current day PnP interfaces are IEEE 1394 (FireWire), Universal Serial Bus (USB), PC card (PCMCIA), and PCI including its variants such as Mini PCI, PCI Express, etc.

2 PnP Architectures

2.1 PnP Requirements in IoT

As mentioned in the chapter's introduction, the vast variety of non-standardized IoT modules has resulted in an enormous heterogeneity that hinders communication and

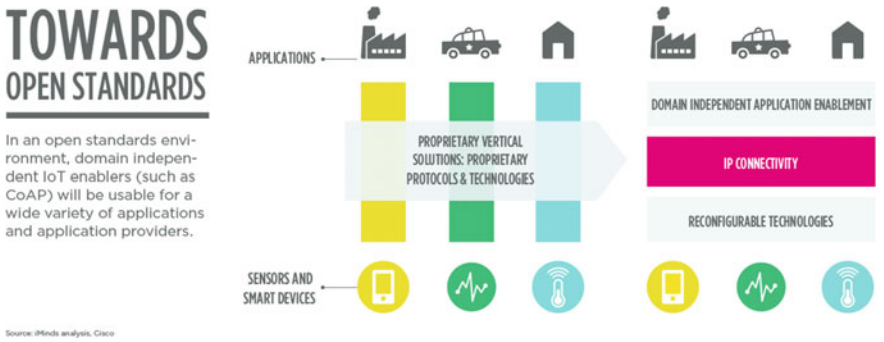


Fig. 1 Towards open standards

interoperability among objects and architectures. Currently, the mainstream practice in IoT development is based on vertical and proprietary solutions, in the sense that a hierarchical bottom to top design (hardware–software-communication) is carried out to suit the needs of each application. This model has to be transformed (Fig. 1) to an open and horizontal structure that will act as a vessel for universal, interoperable, low-cost, and innovative solutions. IoT-enabling tools and services should be developed in a layered and easy to interconnect manner. Following this new model, a multitude of well defined and open IoT-enabling technologies will become available to anyone interested in the IoT sector.

PnP architectures are being developed with the aim of automating or reducing the complexity of the task to configure a new "thing". Configuration, here refers to the procedure which starts when a "thing" is physically connected to a system, to the point that it is able to connect and interact in a network consisting of more "things". The interaction within the network can be either human or machine initiated. A "thing" can be a sensor, an actuator, a communication module, or in general an embedded system with a unique identifier that can communicate in a network either standalone or through some host gateway.

Due to a number of considerations such as cost, size, and available power, things that are used in IoT projects differ significantly from modern day computing systems in terms of resources. Even though there is a multitude of controller units available in the market and one can choose the unit that best fits the application, Table 1 shows key characteristics of four controller units that are predominantly used in different types of applications. As it is expected, a higher amount of resources is available on more expensive and power hungry systems. To achieve a good balance between cost and the set of features that an IoT solution will offer, it is necessary to develop an efficient and effective IoT architecture.

This resource-limited nature of embedded IoT projects, imposes a number of constraints and challenges on the hardware and software schemes that are required for achieving PnP capabilities. Therefore, every IoT-PnP architecture needs to be "lightweight" in a series of aspects. The most important of these aspects are:

Table 1 Key parameters of popular IoT controller units [6–9]

MCU / Resource	ATmega328P	CC3200	SAM9G25	Raspberry Pi 3
Computation				
Architecture	8-bit AVR	32-bit ARM-M4	32-bit ARM A-5	64-bit ARM A-53
Max. frequency (MHz)	20	80	400	1200
Floating-point unit	–	–	–	Yes
MIPS	1/MHz	1.25/MHz	1.57/MHz	2.3/MHz
Connectivity				
Wired	–	–	USB/Ethernet	USB/Ethernet
Wireless	–	Wi-Fi/Bluetooth	–	Wi-Fi/Bluetooth
Memory				
FLASH/EEPROM	32 kB/1 kB	64kB + SD card	64MB	SD Card up to 64GB
RAM	2 kB	256 kB	32 kB on chip	1GB
Max. power consumption (W)	0.06	0.9	0.4	2
Price	2.1 $	12.1 $	7.8 $	35 $

- *Power consumption*: Numerous IoT projects are destined for battery-powered applications. To decrease battery requirements and operational time, the total power consumption of a PnP architecture has to be minimized.
- *Cost*: Cheap IoT modules have been the backbone of the industry's rapid growth. The hardware and software implementation of the architectures should not impose excessive additional cost.
- *CPU overhead*: CPUs that are selected in embedded environments usually have limited computation capabilities. The PnP service should not inflict major CPU overheads that will burden the CPUs typical operations.
- *Memory footprint*: Both program and data memories are constrained. PnP protocols have to be simple and effective.
- *Communication footprint*: In many IoT projects, excessive amount of header or information related to the PnP architecture will result in increased costs (e.g., GPRS data plans.
- *Implementation complexity*: The PnP architecture has to be easily implemented by both, hardware and software engineers.

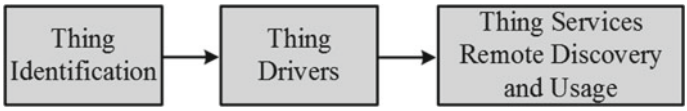


Fig. 2 The PnP process

3 PnP General Architecture

PnP can be defined as a three-step process as shown in Fig. 2. First, the newly inserted to the system thing, is uniquely identified. This preconditions that an identification mechanism is present, and it is compatible among the thing and the system. After identification, appropriate software drivers are required to interface the system with the thing. These are low-level drivers that act upon the control registers of the hardware interface to which the thing is physically attached. Last, users of the IoT network should be able to discover and use the services that are offered by the connected things in order to develop high-level applications. This process corresponds to the process called remote service discovery and usage. In the following sections, each of the above steps will be described in detail.

3.1 Thing Identification

As mentioned before, the identification of new peripherals is the first step for their integration into a system. In mainstream computing systems, such as contemporary desktops and laptops, peripheral identification is realized through custom Integrated Circuits (ICs) that are required for the operation of the interface. These ICs contains data for device identification in special purpose registers. For example, identification in the PCI bus is performed with the use of a file called *Extended System Configuration Data (ECSD)*. This file contains information about the installed PnP devices such as identification data, configuration parameters, etc. This information is read by the BIOS and through the Operating System (OS) system handlers, the identification process gets completed.

However, this approach can only be applied to resource-rich systems and not for embedded IoT devices as it is optimized for performance and neglects CPU overhead, memory footprint and/or power consumption.

Common hardware interconnects in embedded systems for interfacing things and systems are:

- Analog (voltage and current). This interconnection is used for things with analog output. In this case, an Analog-to-Digital Converter (ADC) is required to convert the analog signal to a digitized form is easily processed by the controller.
- Serial Peripheral Interface (SPI) [10]. SPI is a synchronous bus (sampled at a specific clock rate) consisting of four lines: Master In–Slave Out (*MISO*), Master

Out–Slave In (*MOSI*), Clock (*CLK*), and Slave Select (*SS*). Usually, input and output pulses are positive and negative edge sampled which means that in some cases, MISO and MOSI can be connected to the same line. The typical implementation of the SPI interface uses 8-bit messages.

- Universal Asynchronous Receiver Transmitter (UART) [10] which can be half or full duplex. As the name reveals, this is an asynchronous communication scheme between two entities. Due to the asynchronous operation, single or double buffers are required. Furthermore, the communication bit rate (baudrate) and data frame have to be configured manually to match between the transmitter and the receiver. Usually, UART is the TTL/CMOS voltage level application of the RS232/RS485 protocols which are more common in industrial and/or long communication lines. In case of half-duplex operation with no hardware flow control, UART uses only two communication lines T_x, R_x.
- Inter-Integrated Circuit (I^2C) and System Management (SM) Bus [11]. These buses feature a 7-bit addressing scheme making it possible for a total of 128 devices to operate on the same bus. A single master is allowed, which is usually a controller unit and is capable of initiating transactions with the slaves by broadcasting their address. Both, (I^2C) and SM buses require two lines (Serial Clock—SCL and SDA—Serial Data) but differ in the amount of commands that they support; the SM bus command protocol is a *subset* of the commands available in the (I^2C) bus.
- General Purpose Input/Output pins. A number of sensors interact with the controller unit through simple protocols that do not require hardware acceleration and can be implemented only in software (bit banging). Furthermore, GPIOs can be used by sensors to generate main CPU interrupts and/or to control other devices.

Although the above hardware interfaces and corresponding protocols offer simple, lightweight and in some cases fast communication, they lack device identifiers that are essential to a PnP architecture.

As the vast majority of today's sensors are manufactured using one of the above interfaces, a type of middleware which will provide an identification procedure is mandated. After the identification is complete, a set of multiplexers has to switch the peripheral to the appropriate interface bus in order to communicate with the host CPU(s).

3.2 Thing Drivers

After the identification of a thing, the corresponding drivers need to be installed and automatically configured. From a general perspective, there are three ways in which this can be achieved.

First, the thing drivers can be potentially stored in a nonvolatile memory on the thing itself. This is an approach that has been dismissed already by PnP technologies intended for peripheral integration in mainstream computing machines. The reasons were the additional memory required to store the drivers resulting in increased cost

and the complexity deriving from the fact that the machine has to communicate with the device and download the drivers. Furthermore, this is an approach that does not allow for changes to the drivers rendering them outdated for a large span in the device's lifetime. Taking into account, the more cost-sensitive and dynamic environment of IoT projects, such a model cannot be applied during the peripheral integration procedure. The dual approach, i.e., to store the thing drivers on the controller side, is also discarded because of the enormous and continuously increasing variety of available things and hence corresponding drivers.

The current practice during driver development for IoT, is to write driver software after reviewing in detail the thing's datasheet and list of specifications. The developed drivers are written in low-level programming languages in which register manipulation and interrupt handling are required rendering them platform specific. This is a cumbersome task, since the development of a set of drivers that will be both reliable and also efficiently use the thing's functions as a repetitive process. Furthermore, this process leads to nonreusable IoT application code as even trivial hardware changes require updating the driver software.

In modern computing systems, the most common scheme that is followed in PnP peripheral driver integration is the following. Once a new device is connected to and detected by the system, an OS service will start searching local drives and/or remote repositories for the appropriate software drivers. If drivers that match the device id are found, the OS service continues with their installation. After this point, the OS is ready to fully use the peripheral.

In desktop computers, PnP capabilities were first introduced by Microsoft's operating system Windows 95. Microsoft's scheme followed the idea described earlier. The process is illustrated in Fig. 3. When a new device that supports PnP is connected to the system, the service called Plug-and-Play Manager follows a number of steps in order to install the device [12].

- After the detection of a new device, the Plug-and-Play Manager checks the hardware resources required by the device and allocates them.
- The Plug-and-Play manager checks the device's hardware ID and then searches for matching drivers the hard drive(s), floppy drive(s), CD-ROM drive(s) and finally, the Windows Update website.
- Further identification features such as driver signatures or the closest compatible hardware ID used in case multiple drivers are found.
- After security and quality checks, the Plug-and-Play Manager installs the selected driver and the OS is then ready to use the device.

While the above scheme is comprehensive and reliable, the identification and validation layers are not optimized for resource-constrained systems. Identification data are stored in formats (e.g., XML) that are optimized for performance and reliability rather than minimizing CPU overhead and memory footprint. Similarly, in the transport layer, the communication with the remote repositories is performed using rich protocols such as HTTP over TCP/IP. The implementation of the above in a resource-constrained system would be either impossible or would impose nontriv-

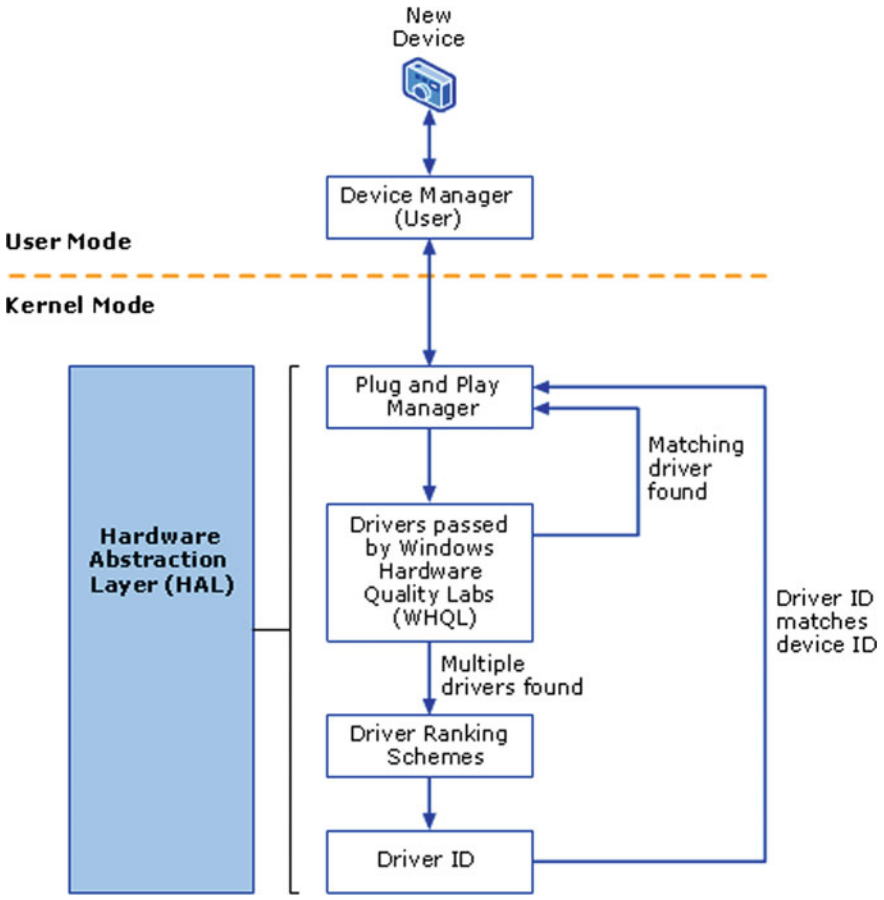


Fig. 3 Microsoft's PnP explained

ial CPU and memory overheads resulting in significant limitations to the range and quality of the developed application.

Thus, it becomes clear that in order to move towards realistic and efficient PnP architectures in the embedded environment of IoT products, simpler and more lightweight protocols that are cross-platform have to be developed while maintaining similar PnP characteristics with resource-rich systems.

3.3 Thing Network Discovery and Operation

After the thing is identified and the software drivers are installed, it is capable of data transactions with its host system. However, the end goal is to create a horizontal

PnP IoT architecture where developers will be able to write applications without a comprehensive knowledge of the underlying hardware but only of their services. For this, a scheme where things publish their available services over a network and developers are able to discover and use them is needed. In mainstream computer systems, this is realized through service discovery protocols. These protocols provide mechanisms that allow to:

- Dynamically discover devices and corresponding services.
- Network users to search and browse the available services.
- Advertise service information crucial to users.
- Utilize an available service.

3.3.1 Service Discovery and Usage in Mainstream Computer Systems

Jini [13]

Jini is a distributed system based on the idea of federating groups of users and the resources required by those users. In a simpler interpretation, the Jini architecture is constructed by a number of hardware and software components that are the infrastructure of distributed system; the Jini registrar provides unicast or multicast detection of this infrastructure (services) by the client returning a proxy object to the clients. This object besides pointers to services can also store Java-based program code that will make use of the service easier to a client. The main mechanism responsible for the communication between services and clients is called *lookup service*. Each Jini device is assumed to have a Java Virtual Machine (JVM) [14] running on it. Jini's main advantage is that it allows users to connect with services without previous knowledge of their address through the lookup service. On the other hand, the existence of the lookup service to manage the interactions between clients and services renders the architecture not suitable for large networks.

Universal Plug-and-Play (UPnP) [15]

Universal Plug-and-Play (UPnP) is a media-independent networking scheme leveraging TCP/IP and other established web protocols. It is developed by an industry consortium called UPnP Forum, which has been founded and lead by Microsoft. Taking into account, the previous discussion about software drivers, UPnP extends Microsoft Windows Plug-and-Play to devices that can communicate in a network. Every device can dynamically join a network, obtain an IP address, announce its services, and also communicates with other devices and services in the network using multicast communication. UPnP is applicable on networks that run Internet Protocol (IP) and on top, it utilizes protocols such as HTTP, SOAP, and XML [16] to accommodate the interactions between the devices. UPnPs main difference with Jini is that it can operate in a decentralized way in the sense that discovery and service advertisement are modeled as events, and are transmitted as HTTP messages over multicast User Datagram Protocol (UDP) [17]. However, this makes the network chatty consuming a significant amount of resources.

Service Location Protocol (SLP) [18]

Another protocol allowing clients to find services over a network without any prior configuration is the Service Location Protocol (SLP). SLP, as UPnP, uses multicast routing in a decentralized way, at least in smaller networks. Each Service Agent (SA)—in our case a thing—multicasts advertisement messages periodically which contain a URL that is used to describe and locate the service. In scaled-up networks, Directory Agents (DAs) exists and cache the information announced by the SAs. User Agents (UAs) can discover services by either multicast requests to the DAs or can listen directly to the announced messages in the absence of DAs.

All of the above architectures have worked well for mainstream computer networks but it is a burden for the proliferation of IoT systems as they are not resource optimized. Jini requires a full-fledged Java Virtual Machine, UPnP operates on verbose XML data representations and SLP mandates further filters due to the lack of device identifiers.

3.3.2 Service Discovery and Usage Protocols in IoT

The Open Connectivity Foundation—IoTivity

Due to the vast expansion of IoT, various company groups have started to develop service discovery and connectivity protocols designed as IoT enablers. One of the biggest, is the Open Connectivity Foundation (OCF) [19] counting more than 300 member companies where among them are Samsung Electronics, Intel, and Microsoft. Its purpose is to "accomodate the communication of billions of connected devices (phones, computers, and sensors) regardless of manufacturer, operating system, chipset, or physical layer". IoTivity [20] is an open multi-layer framework for IoT networks hosted by the Linux Foundation and acts as the reference project implementing OCFs specifications.

Constrained Application Protocol (CoAP) [21]

CoAP, as the name unveils, is an application layer protocol designed for devices with constrained resources in low-power and lossy networks. Typical applications, as defined by the protocol, are wireless nodes that run on 8-bit microcontrollers with low RAM and ROM capacities and slow, high packet error rate networks such as IPv6 over Low-Power Wireless Personal Area Networks (6LoWPANs) [22]. The CoAP design is optimized for Machine-to-Machine Communications (M2M). A very important aspect of CoAP is the easy mapping with HTTP which makes web integration much simpler.

CoAP has been mainly developed and standardized by the Internet Engineering Task Force (IETF) Constrained RESTful enviromnmets (CoRE) Working Group and as complete standard was proposed in 2014. The protocol is maintained and updated by IETF working groups to the current day. CoAP provides resource/service discovery and usage as the aforementioned protocols and is based on the REST model: servers/devices make services available under a URL and clients access these resources through commands such as GET, PUT, POST, and DELETE. These command messages are kept as small as possible to support the use of the UDP protocol in the transport layer and to avoid message fragmentation.

4 Plug-and-Play Models for IoT Applications

4.1 MicroPnP

MicroPnP or μPnP [23–26] developed by the research group of iMinds-Distrinet in University of Leuven, Belgium is one the first complete PnP platforms specifically designed for the IoT industry. μPnP presents a holistic Plug-and-Play architecture including custom hardware for peripheral identification and integration, platform agnostic language for device driver development, and a network architecture for automatic thing discovery and usage.

4.1.1 μPnP Thing Identification

The architecture uses a simple hardware solution to map a large space of addresses to various peripherals. The custom-designed PCB of μPnP contains a set of monostable multivibrators that are capable of generating timed pulses whose length depends on a resistor and a capacitor. On the other side, every μPnP compliant peripheral embeds a unique set of resistors which in conjunction with the fixed value capacitors on the μPnP PCB creates a unique train of pulses. Specifically, four short pulses are generated and each of them is mapped to a single byte value and finally, results in a 32-bit address space. This allows for more than 4 million unique device identifiers. All μPnP peripheral identifiers are mapped to an open online global address space. Identifiers become permanent only after software drivers are integrated in the repository. After the inserted peripheral is identified, it is directed to the appropriate communication bus through a multiplexer and is ready to be utilized by downloading the software drivers.

In [27], the authors presented another model for thing identification. In this case, instead of digitizing the length of pulses generated by multivibrators, a voltage divider technique is applied (Fig. 4). Specifically, the identification resistors (R_{id}) of the thing

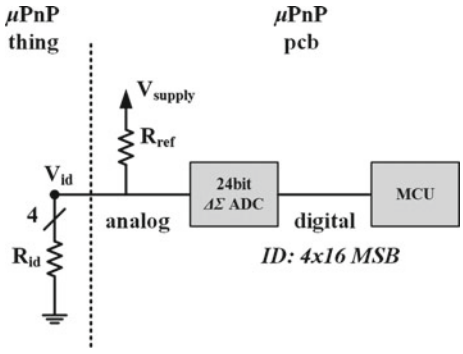


Fig. 4 Identification technique of μPnP v2.0

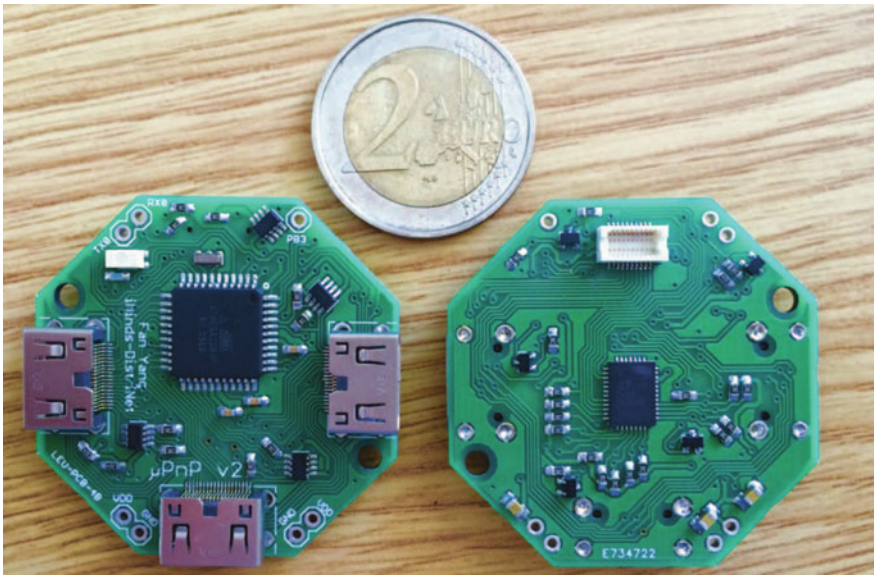


Fig. 5 μPnP peripheral identification hardware v2.0

are connected through a reference resistor (R_{ref}) to the supply rail. An ADC converts the divided voltage (R_{ref}) to a number of bits. With the use of a $\Delta\Sigma$-type ADC, high resolution can be achieved. The authors specify that for 24-bit ADC, the 16 most important bits can be used as a unique tag for things. This creates a 4×16 64-bit address space for thing identification.

Furthermore, both identification schemes require significantly lower energy than a typical USB during the identification process. The implementation of the μPnP v2.0 controller board is shown in Fig. 5.

4.1.2 μPnP Thing Drivers

To achieve multi-platform driver development, μPnP has developed a high-level domain-specific language (DSL). The run-time environment links the DSL with native hardware libraries to the underlying physical interconnects such as ADC, SPI, I2C, etc. The language is event based in order to accommodate the interrupt-driven nature of IoT software.

Platform independence is achieved by compiling the DSL into bytecode instructions that can be interpreted by the μPnP execution environment. This technique is inspired by and similar to Java virtual machines. The various abstraction layers of μPnP's run-time environment are shown in Fig. 6. Five separate elements can be distinguished. The *peripheral controller* interfaces with the μPnP identification PCB and identifies the inserted peripheral. The *driver manager* communicates with the

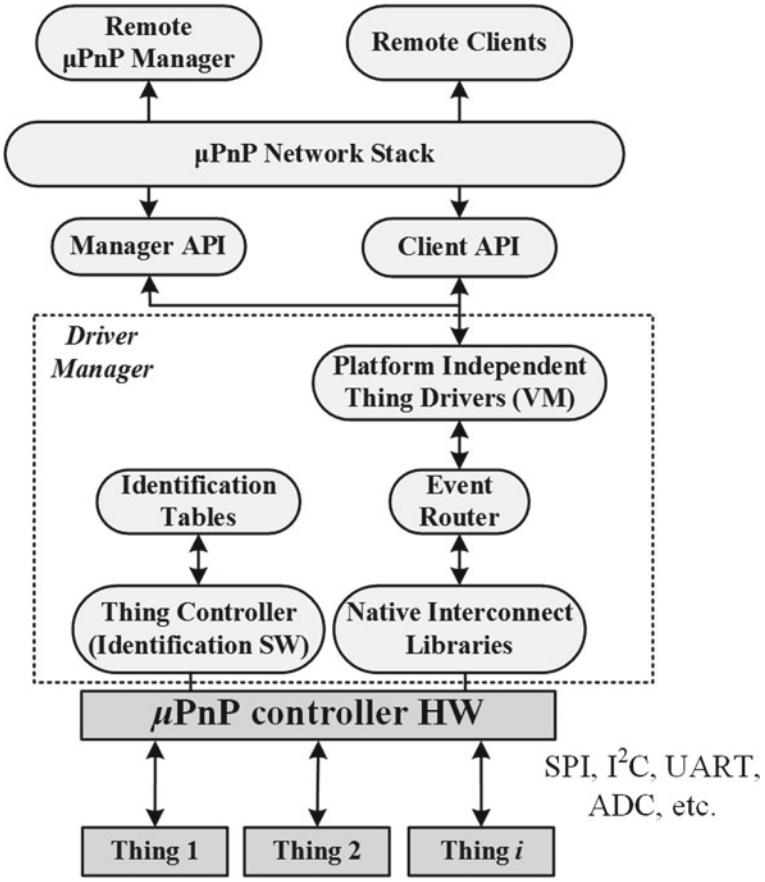


Fig. 6 μPnP execution environment

peripheral controller and gets informed about the connected things and the drivers that are installed. Furthermore, it is responsible for searching and installing available drivers for the newly identified things. A stack-based *virtual machine* executes the bytecode instructions derived from the interpretation of the DSL. Hardware-specific *native libraries* implements the low-level communication with the things. Finally, the *event router* routes the events coming from the DSL, the native libraries, and the software stack. The events are handled asynchronously and do not collide. This is achieved with the application of a FIFO queue for regular events, and a priority queue for error messages.

The thing drivers are installed in the μPnP thing through an entity called Driver Manager. Once the thing is identified, the driver installation process starts with a *driver installation request* towards the anycast address of μPnP server. If a software driver that matches the thing ID is found, the drivers gets downloaded to the μPnP

system. Furthermore, as described in the previous paragraph, the Driver Manager is allowed to query the connected μPnP things about the installed drivers. This is realized by a *driver discovery* message from the host to the thing which awaits a *driver advertisement* message from the thing to the host. The manager is also allowed to remove software driver from a thing using a *driver removal* message. The removal is complete when the thing responds with *driver removal acknowledgment* message.

4.1.3 μPnP Service Discovery and Network Architecture

μPnPs networking scheme features automatic and remote thing discovery and usage like protocols is described earlier and are used in mainstream computer systems. The architecture consists of mainly three software entities. The first entity is the μPnP Thing software which runs on the resource-constrained local machine and allows for automatic peripheral identification. Second, is the μPnP client software which can run on an embedded device or a standard platform and realizes the remote discovery and usage of peripherals that may exist on any node of the network. Last, is the μPnP Manager which runs on server-class machine and is responsible for the remote dispatch and deployment of the thing drivers.

To leverage existing network technologies such as Ethernet and Wi-Fi, the entities are interacting on the network layer through IPv6 over UDP. The thing discovery and usage is realized with three types of communication.

- Unsolicited peripheral advertisements. This source of this announcement is the thing's unicast IPv6 address and contains a set of fields describing it. The unsolicited peripheral advertisements are multicasted to the set of μPnP clients every time a new thing becomes available.
- Peripheral discovery messages. These messages are issued by the clients containing the type of peripheral that is searched. The destination of the messages is the multicast address of all the μPnP things with the specific type of peripheral.
- Solicited peripheral advertisements. These messages are sent in response to peripheral discovery messages. They contain the same information as the unsolicited advertisements but are destined to the unicast address of inquiring μPnP client.

μPnP clients allowed two types of interactions with the things for data production: (a) single value reading and (b) data streaming from the service provider to the client. Writing data to a thing is also supported in case the thing has actuator capabilities.

The above transactions are realized with the following messages:

- *Read* allows a μPnP client to read single value from a peripheral. The μPnP thing that the peripheral is connected to responds with a *data* message which contains the result.
- *Stream* messages are send from clients to things when a client wants to subscribe to a continuous stream of data. In this case, the thing responds with an *established* message which contains the multicast address that the client should join. *data* messages are send continuously from the stream address and a *closed* message is broadcasted if the stream stops to inform all the μPnP clients.

- The control of peripheral is achieved with a *write* message. This message is sent from a client to a thing. If the write process is completed successfully, an *acknowledgment* message is given as response to the client.

4.2 IEEE 1451 Standard [28]

The IEEE1451 is a family of "smart transducer" interface standards developed by the Institute of Electrical and Electronics Engineers (IEEE) and was first released on 1997. A smart transducer is defined as the integration of an analog/digital sensor or actuator, a processing unit, and a communication interface. According to this definition in Fig. 7a, we can see the structure of a smart transducer. It consists of (1) sensors/actuators, (2) signal conditioning and/or data conversion circuits, (3) host processor, and (4) network communication. The communication paths in this structure are two-way as data can flow from the transducer to the network in the case of a sensor or from the network to the transducer in the case of an actuator.

An IEEE1451 smart transducer should have features like self-identification, self-description, self-diagnosis, self-calibration, location-awareness, time-awareness, data processing, standard-based data formats, and communication protocols. The IEEE1451 aims to achieve the above with the design and integration of *Transducer electronic data sheets or TEDS*. The architecture is shown in Fig. 7b. The modules that constitute the architecture are: (1) a Network Capable Application Processor

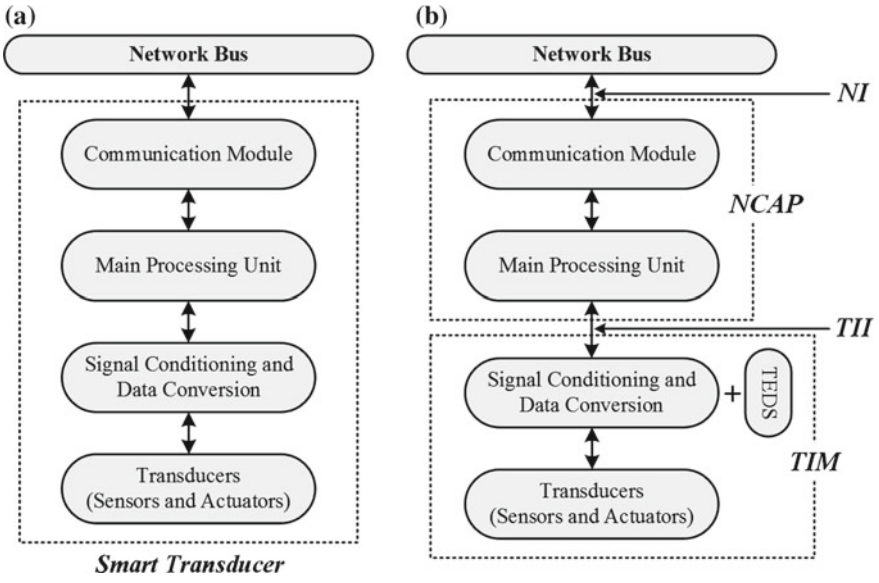


Fig. 7 **a** A smart transducer model; **b** This architecture adds TEDS and the system partition into

(NCAP), (2) Transducer Interface Module (TIM), and a (3) Transducer-Independent Interface (TII) for communication between (1) and (2). The TII is defined by a communication medium and a data transfer protocol with messages like read, write, read, and write, etc. The smart transducer can be connected to the network through any common Network Interface (NI).

The key feature of IEEE1451 is the TEDS. The TEDS can be stored in a non-volatile memory space attached to the Smart Transducer Interface Module (STIM) containing necessary information for the host system to interface with the transducer, such as identification, calibration, and correction data. Also, a virtual TEDS can be implemented, allowing legacy sensors and transducers without any storage space to be included in the standard. Four kinds of TEDS are *mandatory* for the application of the standard, and are as follows:

- Meta TEDS.
- TransducerChannel TEDS.
- PHY TEDS.
- UserTransducerName TEDS.

Optional TEDS includes Calibration TEDS, TransferFunction TEDS, Location and title TEDS, and Frequency Response TEDS.

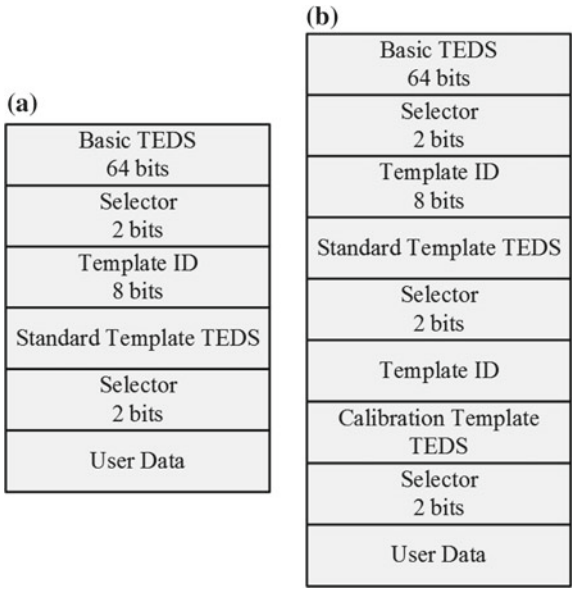
Since 1997, the standard has developed and improved in order to accommodate modern miniature sensors and actuators as well as different communication/network protocols. The full list of released protocols is:

- 1451.0-2007 Common Functions, Communication Protocols, and Transducer Electronic Data Sheet (TEDS) Formats
- 1451.1-1999—Network Capable Application Processor Information Model [29]
- 1451.2-1997—Transducer to Microprocessor Communication Protocols and TEDS Formats [30]
- 1451.3-2003—Digital Communication and TEDS Formats for Distributed Multidrop Systems [31]
- 1451.4-2004—Mixed-Mode Communication Protocols and TEDS Formats [32]
- 1451.5-2007—Wireless Communication Protocols and Transducer Electronic Data Sheet (TEDS) Formats [33]
- 1451.7-2010—Transducers to Radio-Frequency Identification (RFID) Systems Communication Protocols and Transducer Electronic Data Sheet Formats [34].

4.3 The TEDS Structure

The TEDS are encoded using specific templates to maintain a balance between the provided transducer information and the amount of memory that needs to be occupied. In the IEEE1451.4 standard, the TEDS is defined as multisection template. These sections are chained together to form a complete TEDS (Fig. 8). The first

Fig. 8 TED examples



section is the basic TEDS and contains essential identification information. Depending on the application and the complexity of the transducer, further sections can be followed. The type of the following section is indicated by 2-bit selectors. The last section, is an open user area where further information of instructions that are not defined in the template can be given.

Basic TEDS

The basic TEDS has a length of 64 bits. The first 14 bits are reserved for the Manufacturer ID, 15 bits for the Model Number, 5-bit character code for the Version Letter, 6 bits for the version number, and 24 bits for the serial number of the device. The assignment of Manufacturer IDs and other binary data are provided in ASCII files either from IEEE or from the manufacturers.

Several researchers have proposed modular PnP-like models based on the IEEE1451 standard for various applications [35–37]. In [38], the authors have leveraged the the standard in order to create a thermal comfort sensing system for buildings. In [39], a configurable Wireless Sensor Network (WSN) is developed based on the IEEE1451 and a Complex Programmable Logic Device (CPLD). The authors have used the CPLDs high parallel throughput and the interoperability offered by the IEEE1451 standards to build a flexible and reconfigurable water quality monitoring WSN.

In [40], a detailed example of how the IEEE1451 standard was practiced for the needs of an application is presented. The transducer was an electrogoniometer, which is a transducer crucial to physiotherapy applications, is designed based on the IEEE1451 standard. Specifically, IEEE1451.2 and IEEE1451.0 were implemented.

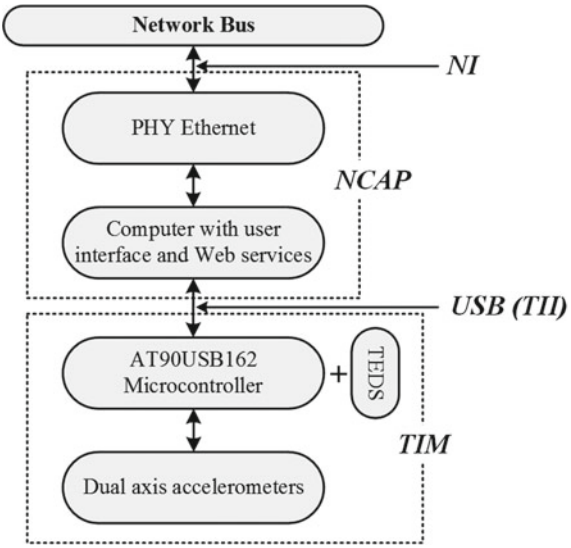


Fig. 9 Block diagram of IEEE1451 compliant electrogoniometer

The block diagram illustrating the implementation of the sensor is shown in Fig. 9. The STIM consists of two dual-axis accelerometers as sensing elements and an Atmel AT90USB162 microcontroller. The embedded Microcontroller Unit (MCU) is programmed as a STIM using C and the TEDS is stored into its nonvolatile memory (FLASH). Furthermore, it features integrated full-speed USB peripheral for communication between the STIM and the NCAP. IEEE1451.2 is implemented on top of the USB interface to render the sensor IEEE1451 standard compliant.

Specifically, the interfacing sequence can be described as the following. Once the STIM is connected to the NCAP, the STIM sends a *Tim Initiated Message* to annouce its existence. Next, the PHY-TEDS information is announced to the NCAP using the Publish–Subsribe method. Several values such as, TIM identification, Communication Module ID-Type-Name-Object, STIM Channel numbers-ID-name, etc. Furthermore, the NCAP handed the TEDS information including Meta TEDS, TransducerChannel TEDS, User's Transducer Name TEDS, Manufacturer ID, Version of TEDS, Number of Channel, and Serial Number. Finally, the NCAP send the commnad to acquire the sensor data.

The NCAP in this case is developed with the Java Development Kit and the Eclipse IDE, and it operates on a standard commercial laptop computer.

Reconfigurable Wireless Sensor Networks (WSNs)

In [41], the authors present a reconfigurable WSN with PnP capabilities regarding the network architecture of the testbed. The proposed architecture allows automatic configuration (Plug) of the network by utilizing a Zeroconf protocol that sets up a multi-hop network. Furthermore, reconfiguration and experimentation (Play) is achieved on the basis of RESTful interactions with each node. The node is composed

of configurable transceiver(s), configurable/modular protocol stack, and a monitoring/control module.

The configurable transceiver has to be low cost/power and support parameter reconfiguration. In the author's implementation, two transceivers were used for dual-band communication. The first one was the reconfigurable TI CC1101 operating at the 868 MHz RF band and the AT86RF231 operating at 2.4GHz. The latter, is 802.15.5 compliant which means it is suitable for low-power 6LoWPAN communication.

A configurable/modular protocol stack is selected for efficient development and experimentation on existing protocols. Furthermore, the configurable/modular stack has to reconfigure the management network making each device a uniquely addressable and accessible network thing. The configurable/modular platfrom used was the CRime stack [42] and Contiki [43] 6LowPan/IPV6 stack was used for the management network. This management network stack is based on the Routing Protocol for Low-power and Lossy networks (RPL) [44]. RPL is a protocol for automatic network discovery and configuration. A dual-stack Contiki implementation is leveraged to run both protocols in parallel.

The monitoring and control module operates using CoAP which is an HTTP-like protocol redesigned for devices with constrained resources, as described in previous sections. A set of CoAP handlers enables system users to remotely configure and operate the testbed (*Play*). Specifically, the CoAP over UDP handlers (messages) that were developed allows the user to perform the following interactions:

- Remote monitoring and diagnosis of the system.
- Remote parameter tuning.
- Over the air software updates and upgrades.

Reconfigurable Wireless Sensor Nodes

Mikhaylov and Huttunen [45], Mikhaylov and Paatelma [46] introduces a type of PnP nodes for a Wireless Sensor and Actuator Network (WSAN). The concept that the authors want to achieve in this work is the fully modular development of a WSAN using modules which can virtually be anything; ranging from power sources, wireless/wired communication hardware, and controller units to sensors, actuators, encryption devices, localization hardware and/or additional memory. After physical connection of the available modules, the Main Processing Unit (MPU) should be able to identify all attached modules and download the necessary drivers using local or network connection.

The proposed hardware architecture for achieving automatic peripheral identification and integration is presented in Fig. 10. A new interface called Intelligent Modular Periphery (IMP) Interface (IMPI) is defined and is responsible for interfacing the peripherals to the MPU. The IMPI can be disseminated in functional terms in (a) the power supply lines which consist of the input voltage, output voltage and ground (Vin, Vout, GND), (b) the IMP bus lines that are reserved for device identification and control, and (c) underlying interface buses that the peripherals may require.

A Module Control and Identification Unit (MCIU) stores all the required information about the module as well as about the peripherals hosted by the module. The

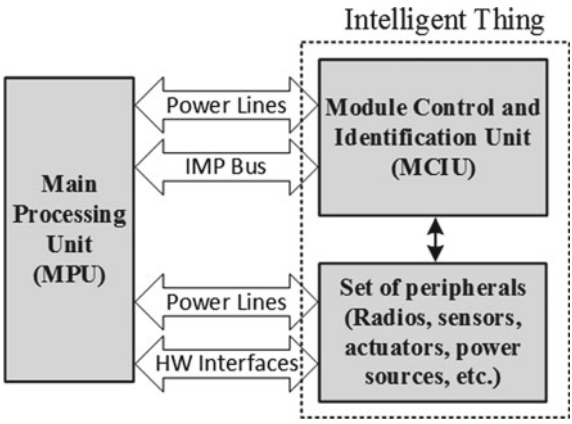


Fig. 10 Hardware architecture

MCIU is accessed through the IMP bus and can also provide rudimentary control over the peripherals such as power management. Physically, the MCIU can be a microcontroller, a PLD, or any other logic device with similar functionality.

The IMP bus is implemented by daisy chaining the well known and ubiquitous SPI bus. Through the IMP bus, the MPU first discovers the total number of connected peripherals. Then, it downloads the peripheral description data (PDD), which consists of the Peripheral Connection Descriptor (PCD) and the Peripheral Service Data (PSD). The PCD contains the required data for the MPU to map the particular communication interfaces to specific modules and peripherals. The PSD on the other hand, contains information such as name, identifier, SW drivers, calibration coefficients, etc. Using these information, the MPU can automatically discover and use the attached peripherals.

On the software side, the dynamic underlying hardware requires a complex architecture in order to become fully functional and efficient. The proposed architecture is presented in Fig. 11. The main component of this middleware is called Resource Manager and has three major building blocks. The first one, is the Module manager, being responsible for low-level operations such as identification of peripherals and modules and interrupt prioritization. The second building block is the Communications manager and its task is to handle the communication of the node with the network. This manager asserts the existing communication transceivers and decides which must be used and under which parameters. Among the responsibilities of the Communications managers are also discovery of devices and networks, mapping/translating of network addresses, etc. Last, the Applications manager supervises and controls the launch/stop of every available application or service. The Applications manager also acts as a broadcaster of the node services in a network scheme.

The PnP Web Tag

As it was mentioned in previous sections, the current required knowledge for someone to develop a full-stack IoT application is low-level embedded programming, networking mainly using low-power protocols and transceivers and web integration.

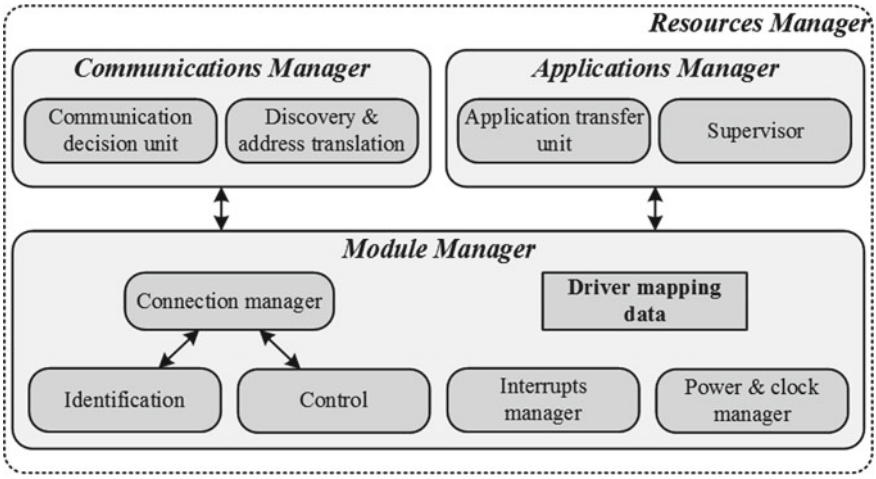


Fig. 11 Software architecture

These multidisciplinary skills hinder a lot of individuals and companies with low resources from entering the industry. In [47], the authors aim to address this problem by developing a PnP programming model for connecting IoT devices to the web.

The tool that is developed has many cores:

- A parser, that identifies IoT services in HTML pages. The parser is implemented as a pure client-side JavaScript. The parser enables to support the new PnP web tag.
- An HTML instrumenter, that refreshes in a dynamic way the HTML page as new data arrives. The instrumenter is configured by the parser and is also realized in pure-client JavaScript. Furthermore, it sends commands to the IoT devices from the JavaScript applications.
- A proxy server that acts as the interconnection between the instrumenter and the CoAP IoT services. It is implemnted in the IoT network gateway, and it bridges the CoAP protocol and the WebSockets protocol that is used by the HTML and JavaScript elements.
- JavaScript allows the web developers to ignore the low-level embedded programming and build complex sensing and control software.

PnP transducers in Cyber-Physical Systems

However, in modern IoT networks which may contain thousands of things, the added cost and the increased hardware complexity that is imposed by the standard, poses a major burden to its proliferation and widespread establishment.

Specifically, the standard defines 16 TEDS templates. The number of transducers available is increasing exponentially over the last years. A new template, requiring by the user full knowledge of the complex standard has to submitted for a new transducer to be included in the protocol. Furthermore, although IEEE1451 is an

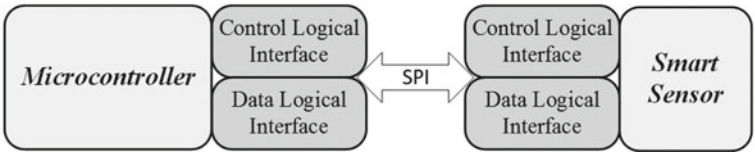


Fig. 12 Logical interfaces in the Plug-and-Play architecture

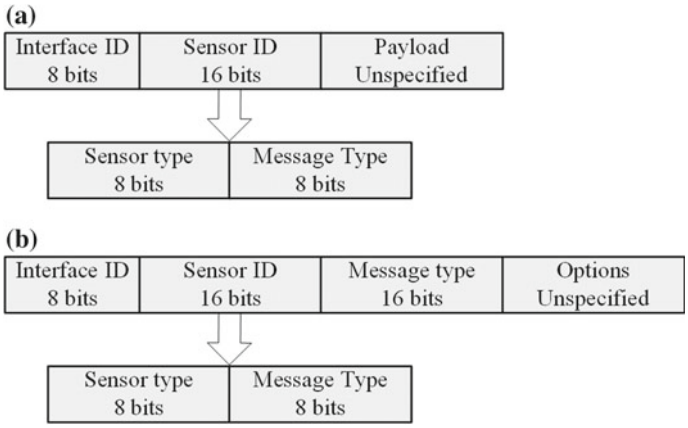


Fig. 13 **a** Data frame structure, **b** Control frame structure

open standard, a large number of these transducers contain proprietary information leading to proprietary TEDS. Second, the standard includes byte-oriented messages, thus requiring significant amount of memory for resource-constrained devices and causing nontrivial CPU overheard.

Taking into account the above, the authors in [48], proposed a new transducer/interface independent and lightweight method for PnP transducers based on the SPI bus. Over the physical SPI interface, two virtual interfaces are defined: a control and a data one Fig. 12. The physical lines are distinguished with the use of headers in the byte frames during communication. As shown in Fig. 13, the control frame communication structure includes a 2-byte message type to distinguish it from the data frame structure. With this 2-byte representation 65536 different messages can be generated, a number that can support a vast majority of applications.Please provide better clarity in the sentence “With this ... applications” and amend if ncecessary.

The identification of the modules from the main processor is implemented in three steps. First, the controller sends a Description Request message to each available slave bus. After sending the message, the controller expects a Description response message. If this message is received, the master controller records the specific slave SPI port as occupied and transmits a Description ACK message.

Plug-and-Play Software Components in Industrial Cyber-Physical Systems

In [49], the authors are using a service-oriented software architecture in order to achieve PnP in industrial systems. Specifically, a paradigm transformation is targeted,

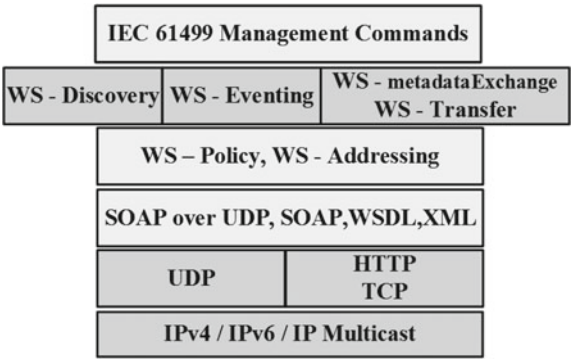


Fig. 14 Device profile for web services for Plug-and-Play software services in iCPS

from the current practice where software engineers have to develop the required software for each device and then integrate it to the system, to a model that software engineers can use directly services that are published from the devices to create high-level applications ready for system integration.

To implement such an architecture, the first step is to define the protocols and operating blocks for system level modeling. The authors propose IEC 61499 function blocks [50] for the system level modeling and Web Service Description Language (WSDL) [51] for the interface protocols.

WDSL is a language for accessing web services and is based on the Extensible Markup Language (XML). A WDSL document can define a number of elements required for service providers and users to communicate. Such elements are data types, messages, portType, binding, and services.

The above are used to render the available functions of the installed devices as callable services. For automatic service discovery, the authors select a WS-discovery type protocol which is called Device Profile for Web Services (DPWS) [52]. In WS-discovery type discovery protocols, the system services are stored dynamically in distributed registries. The complete DPWS protocol stack is shown in Fig. 14. It utilizes WS protocols based on SOAP, WSDL, and XML architectures. UDP or HTTP/TCP over IP is used in the transport layer. To enable PnP, IEC61499 management commands are used in the application layer. The typical messages used for service discovery are Hello, Probe, Bye, Resolve, Put, Get, Create, and Delete.

Generic Sensing Platform

Finding the right off-the-shelf platforms/devices for an IoT system can be a challenging and time consuming task. However, the design cost would be minimized if reconfigurable and efficient (in terms of power, size, and communication protocol compatibility) generic sensor node/platforms were available for IoT design.

In [53], the authors recognize the need of modularity and interoperability between IoT devices and sensors, and proposed a reconfigurable RFID (Radio-Frequency Identification) sensing tag as a Generic Sensing Platform (GSP) featuring PnP capabilities. RFID is a technology that suits IoT applications well as it offers low-power consumption and small size. On the other hand, RFID tags are significantly constrained in terms of sensing, computing, and data logging capabilities. Moreover,

RFID sensing tags have to be accompanied by an RFID reader to be able to operate and sense.

The authors present an approach for the design of a Gen-2 [54] compatible and semipassive RFID GSP-tag. The GSP-tag is designed to accommodate a variety of sensors, multiple sensing channels, and PnP. Furthermore, the tag can operate in two modes (1) continuous data transmission mode (online) and (2) data logging mode (offline). The first three memory blocks of the user memory have been reserved and act as configuration bytes.

Specifically, the GSP-tag consists of two fundamental building blocks. An analog front-end and a digital core as shown in Fig. 15. The analog front-end contains the 915 Mhz meandering antenna as well as the L-C matching network and the modulation–demodulation circuitry. The analog front-end is power passive, so the operation of the modulation–demodulation depends on the reader's transmitted power. The digital part is implemented using an ARM Cortex-M3 microcontroller and it is battery powered. The MCU is responsible for acquiring the sensed data from the peripheral devices and performing the digital baseband communication through Gen-2 protocols with the RFID reader.

The Serial Shipping Container Code (SSCC-96) [55] EPC standard was used to perform data transmission from different channels and also act as GSPs identification mechanism. The authors as there is no global standard for RFID sensing applications, modified the EPC tag standard to provide identification for the GSP. Fields such Header, Filter, Partition, Company prefix, and Serial No. comprises a typical EPC ID. Among these, the Serial No. represents a secondary identity of the tag. Therefore, this field was modified in order to carry the variable sensor data. With this modification, automatic identification and sensor data transmission as well were achieved.

A Scalable and Self-Configuring Architecture for Service Discovery in the Internet of Things

In [56], the authors recognize the need for an architecture capable of accommodating billions of IoT nodes with minimum human intervention and proposed a Scalable and Self-Configuring Architecture for service discovery. In the paper, it is stressed that a service discovery protocol should enable communication between (1) things that are concentrated and for example, belong at the same subnetwork and (2) things that can operate within a broader scale and multiple subnetworks. Furthermore, such a protocol needs to be scalable taking into account the rapidly increasing number of devices.

The enabler which the authors propose to render the above possible is a Peer-to-Peer (P2P) network with zero-configuration (Zeroconf) mechanisms at the local scale. A dedicated boundary node, called a "IoT Gateway", acts to gather information about the resources of the locally attached nodes and create a Resource Directory (RD). This information is stored and can be accessed by other clients among the P2P network allowing for automatic Service Discovery (SD). Such a server-free approach is scalable and makes the performance of the service discovery to depend only on the size of the IoT network. To avoid application specific constraints, the architecture is designed to consist of format-of-service and resource-descriptor agnostic components.

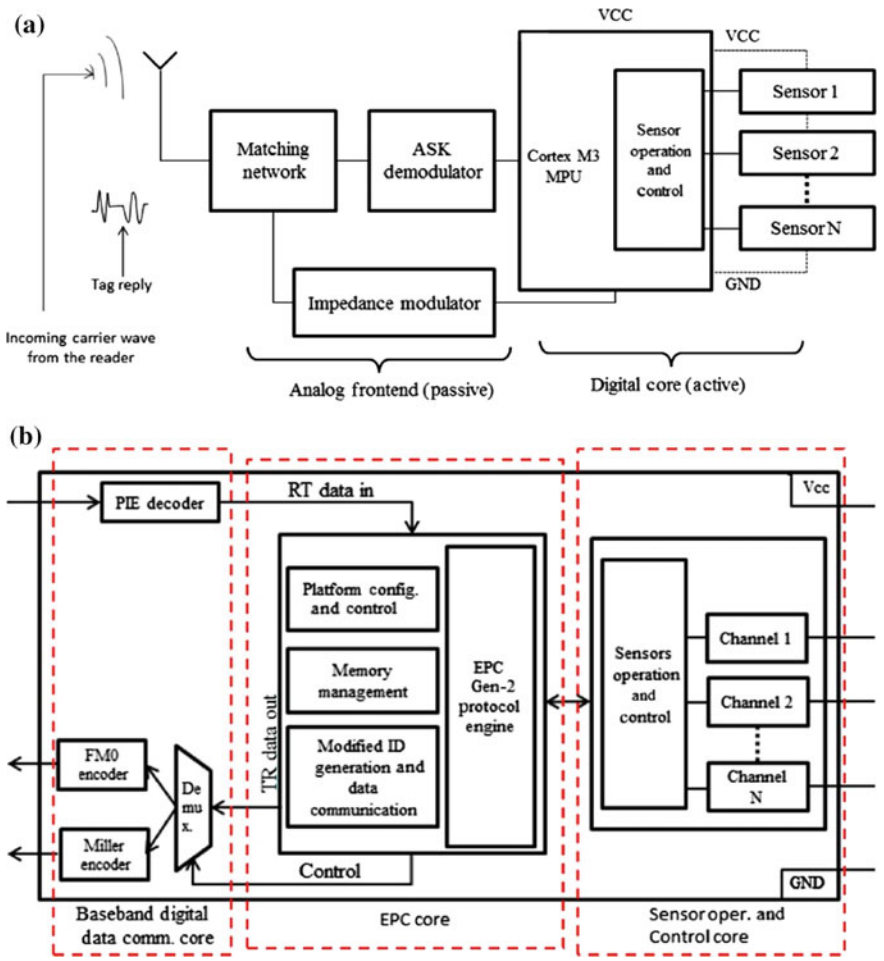


Fig. 15 Implemented GSP-tag on a PCB with off-the-shelf components

The IoT gateway is responsible for providing to the nodes service discovery, caching, proxying, and access control functions. The *Proxy Function* of the gateway is performed on the application layer using CoAP. According to CoAP specifications, the gateway may act as a CoAP origin server and/or proxy. A CoAP endpoint is defined as an origin server when the resource has been created locally at the endpoint. A proxy endpoint, implements both the server and client side of the CoAP, and forwards requests to an origin server and relays back the response to the inquiring node. A proxy may also be capable to perform caching and protocol translation.

An IoT gateway architecture can be distinguished in three elements as shown in Fig. 16. Specifically, the architecture consists of the following:

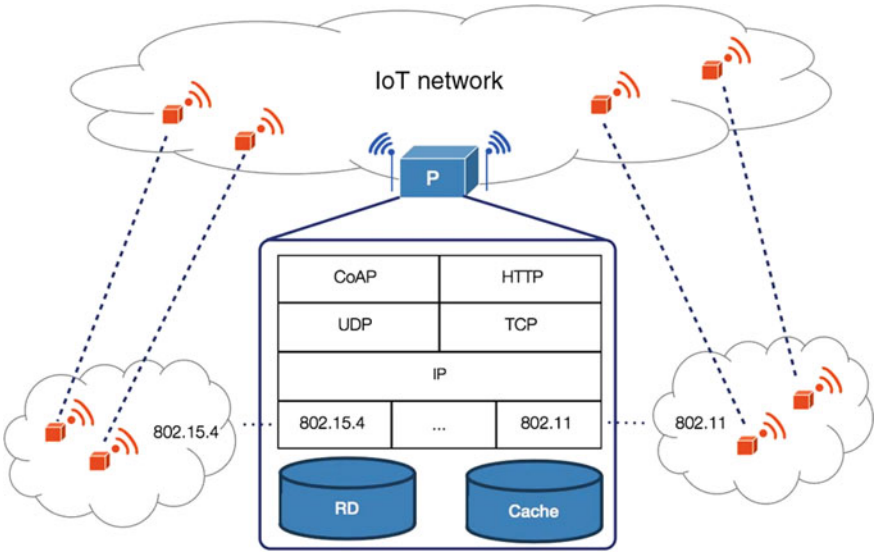


Fig. 16 Implemented GSP-tag on a PCB with off-the-shelf components

- An IP gateway capable of managing IPv4, IPv6 connectivity among things in various networks.
- A CoAP origin server that can be leveraged by the CoAP clients to post resources which are going to be maintained by the server.
- An HTTP-to-CoAP translation service for accessing services and resources are available in an internal constrained network.

In the local environment, Zeroconf is used to automatically detect and configure the new nodes that join the network. The implemented SD protocol supports in general two cases for its application:

- A new thing that joins the network publishes new services.
- A client thing, already existing in the network, discovers available services offered by other connected things.

On the first case, the process followed is different depending on whether the thing is a CoAP origin server or a server client. When the thing is CoAP server it can be queried directly about its services. On the other hand, when it is a CoAP client, it pushes the information about the available services to the IoT gateway, which acts as a RD.

Integrating Transducer Capability to GS1 EPCglobal Identify Layer for IoT Applications

The issues of resolving vast heterogeneity among IoT things is also addressed in [57]. The authors propose the application of the Electronic Product Code (EPC) global Identify Layer and IEEE1451 for building a uniform IoT architecture. The uniformity relies on the representation of the raw data collected from the sensing

elements in standardized format and the PnP capabilities offered by the IEEE1451 standard.

EPCglobal is a set of standards for sharing data within and across organizations while IEEE1451, as discussed previously, is a set of standards for communication betwwen smart transducers and networks. The authors propose the adoption of IEEE1451 Smart Transducer Standard and its integration with an extension of the GS1 EPCglobal architecture.

GS1 EPCglobal architecture is utilized by numerous supply chain management systems to create track and trace applications through RFID tags. The GS1 EPCglobal is a three-layer architecture and consists of the: (1) identify, (2) capture, and (3) share layer. The identify layer gathers identification and self-awareness data. According to the architecture, the tag data coding protocols are defined at Tag Data Standards (TDS) and Tag Data Translation (TDT) which are extendable. The modification is made to the Serialized Global Transducer Item Number (SGXIN) [55] where the TDT file is appended in order to transform it to a *Thing* Data Translation. In this way, the new TDT can accommodate transducer and tag data concurrently. In the capture layer, the raw data acquired from the identify layer are filtered by the Application Level Events (ALE) middleware. The ALE middleware is an application of the corresponding standard and specifies the interface through which end users are able to obtain consolidated data and information about physical processes from a multitude of sources. The extended version of the ALE middleware is capable of handling raw data not only from the RFID tags but also from the smart transducers.

Finally, in the capture layer, the filtered and grouped raw data can act as input to capture service and generate specific events. The events are inserted into an extended EPC Information System (IS) which will be able to accept transducer functionalities. Then, the extended EPCIS stores the events and renders them available for query.

The integration of the IEEE1451 compatible transducers to the EPCglobal architecture is the main focus of the paper. Metadata are gathered from IEEE1451 TEDS structure and are appended to the identify layer of the EPCglobal standard (Fig. 17). The authors have distinguished as minimum meta-TEDS the transducer ID, sensing/triggering data, and other data required for handling all the relevant data as a block. These include the Universal Unique Identifier (UUID) which consists of metadata such as location (42 bits), manufacturer (4 bits) year (12 bits), and time (22 bits), summing up to a total of 80 bits.

The PnP nature of IEEE1451 is coupled with the well defined in RFID applications EPCglobal series of protocols to create a flexible framework for developing IoT produces/services.

Sensor Discovery and Configuration Framework for The Internet of Things Paradigm

Perera et al. [58] proposes another platform for automatic sensor discovery and configuration called *SmartLink*. The *SmartLink* architecture is based on a software entity that the authors have named as Context-Aware Dynamic Discovery of Things (CADDOT). The model consists of a total of eight phases that are performed either by the *SmartLink* or a cloud-based IoT middleware.

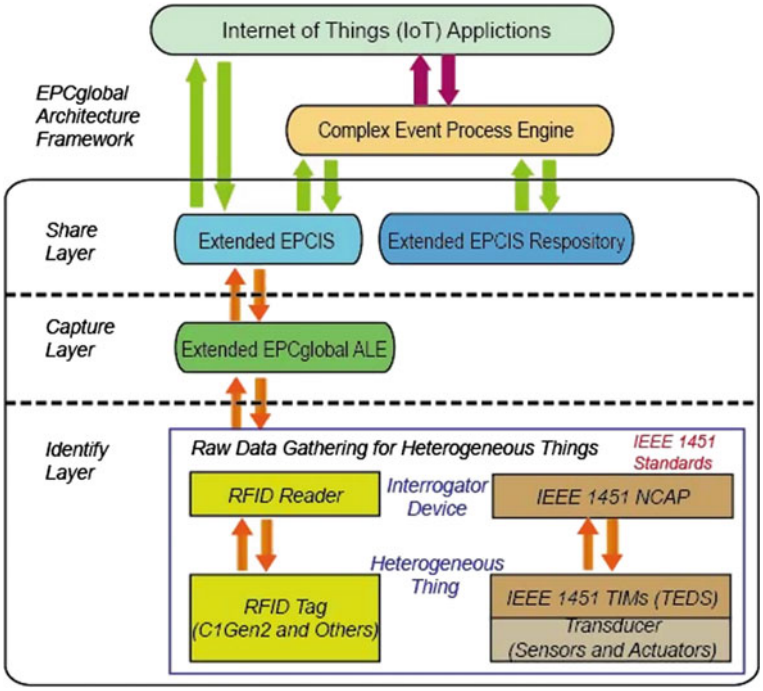


Fig. 17 EPCglobal extension with IEEE 1451 capability

The eight steps are:

- *Detect*: Here, the assumption is that the sensors are configured to seek for networks (Wi-Fi/Bluetooth) to connect to without the need of authorization. *SmartLink* is configured to act as an open wireless hotspot so that sensors connect to it in an ad-hoc fashion.
- *Extract*: In this phase to bypass the heterogeneous messages that are needed to identify every sensor, the authors propose to add an extra layer of communication information with which every sensor connected to the ad-hoc network of *SmartLink* will be able to respond to the message "WHO". The response to this message is the minimum amount of information for a sensor to be identified such as the unique identification number, model number/name, and manufacturer. After this response, the *SmartLink* has the required information to proceed with further interactions with the sensor through the sensor's native communication protocol. This approach of identification is similar to the TEDS algorithm mentioned earlier.
- *Identify*: During this phase, *SmartLink* sends response from the newly detected sensor to a cloud-based IoT middleware. The middleware queries its databases and retrieves every data available regarding the sensor. With this procedure, the sensor's profile in identified completely.

- *Find*: After the sensor module is fully identified, the IoT middleware pushed the necessary software drivers from the its database to the *SmartLink* where they are installed.
- *Retrieve*: At this point, *SmartLink* is capable of full communication with the sensor. Using this capability, the sensor is queried on any additional configuration details might contain such as schedules, sampling rates, data structures, etc. Further, *SmartLink*, if possible, it will communicate with other online sources to retrieve additional useful information related to the sensor.
- *Reason*: In this step, a context-aware sensing software is deployed. The IoT middleware takes into account the inputs from multiple sources to evaluate the capabilities, limitations, and operation details of every sensor. Following an optimization process a comprehensive sensing plan for each individual sensor is designed.
- *Configure*: In this last phase, sensors and cloud-based IoT software systems are going through final configurations. Schedules, communication, sampling frequency, and other details that were designed in the previous step are installed on the sensors. Communication between the sensors and the IoT cloud software is established through direct connection or through network capable gateways. Finally, communication configuration such as IP address, port, and authorization are provided by the *SmartLink*.

The authors point out possible application of this architecture to home automation and/or to agriculture IoT. A home automation system based using a Raspberry Pi as *SmartLink* was developed to demonstrate the effectiveness of the scheme.

Agile Manufacturing

Authors in [59] discuss about *Agile Manufacturing* [60] and how IoT in conjunction with PnP models can act as a strong enabler for industrial systems that will be easily configured and installed. Also, the notion of IoT@Work [61, 62] is mentioned, in which large number of intelligent devices are automatically configured in a similar manner to contemporary USB devices.

The discussion is about PnP systems and modules used in industrial premises. After thorough analysis of numerous manufacturing processes and systems, the IoT@Work project has been identified and is aiming to fulfill eight general requirements (A-Gx) towards PnP agile industrial systems:

- A-G1. System modification should be allowed during runtime, if the remaining systems in the production process are not affected.
- A-G2. When modifications require the production system to be in an offline state, rapid re-initialization should be possible.
- A-G3. When the system is modified during run-time, controlled initialization should be realized.
- A-G4. In case of communication faults, automatic network rerouting has to be supported.
- A-G5. Provision of a flexible and independent system bootstrapping.
- A-G6. Device migration and rapid device reconfiguration has to be supported.
- A-G7. Minimization of manual effort for configuration and initialization of the system.

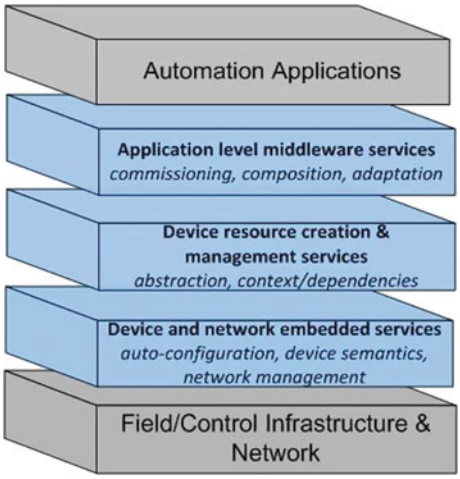


Fig. 18 IoT@Work layered architecture

- A-G8. Intelligent system responses to various events such as faults.

Furthermore, the initiative has identified four additional requirements specific for automotive manufacturing systems.

- A-A1. Provision of a Graphical User Interface (GUI) for network configuration and initialization.
- A-A2. Provision of a GUI for the network maintenance system.
- A-A3. Provision of flexible and reliable semantic addressing scheme.
- A-A4. Capability of remote control and maintenance, in case of external maintenance contractors.

To address the aforementioned requirements, IoT@Work proposed a layered system architecture. Each layer corresponds to a different functional group and the abstraction of the layers begin with the low-level embedded devices and its end point is automation applications. In particular, three functional groups are identified Fig. 18 and can be defined as:

- The lowest abstract layer includes all the devices and network infrastructure along with their management functions. These functions are identifier assignment, device semantic and context collection, communication interfaces configuration, etc.
- The middle layer refers to all the services and functions that become available through the existing infrastructure and installed hardware systems. The abstraction level of these resources are higher since a lot of details from single devices are hidden.
- The top layer of abstraction refers to all the control schemes and scenarios that can be developed to service specific applications. At this layer, the interpretation of the application logic is performed during configuration time and runtime.

An Integrated Device and Service Discovery with UPnP and ONS to Facilitate the Composition of Smart Home Applications

Mitsugi et al.[63] proposes a complete PnP solution oriented towards smart home applications. Specifically, the authors present a protocol with *device* and *service* discovery capabilities by integrating the Universal Plug-and-Play protocol (UPnP) with the Object Naming Service (ONS) [64].

UPnP can keep a list of the available devices and services using its simple service discovery protocol. Every time that a new device joins the local network, it sends a message *ssdp:alive* to the gateways of the network in order to inform its existence. However, the gateway corresponds with the device using XML over HTTP. Such an implementation in many cases might not be possible because of the resource-constrained nature (computation, memory, and networking) of many devices used in a IoT smart home environment.

To bypass this issue, the Object Naming Service (ONS) is integrated to UPnP protocol. The ONS refers to the global service directory based on the EPC-GS1 standard (4.3). A control point (gateway) collects the device identifiers and then retrieves the service through ONS. An ONS client installed locally, can query the Root ONS with an EPC as key and find all services associated with the EPC. By this, services are available to the developer to create smart home applications.

The integration of the EPC protocol with UPnP is possible using CoAP. During an *ssdp:alive* message in a CoAP frame, the URI option is used to determine the device's EPC, which will be its unique identity. The operation of the developed protocol is shown in Fig. 19. To update its list of available service in the network, UPnP has a feature called *m-search*. The authors, again taking into account the resource-constrained devices on which the protocol will be implemented, developed a *transparent m-search* which has been shown to perform better [65] in such circumstances.

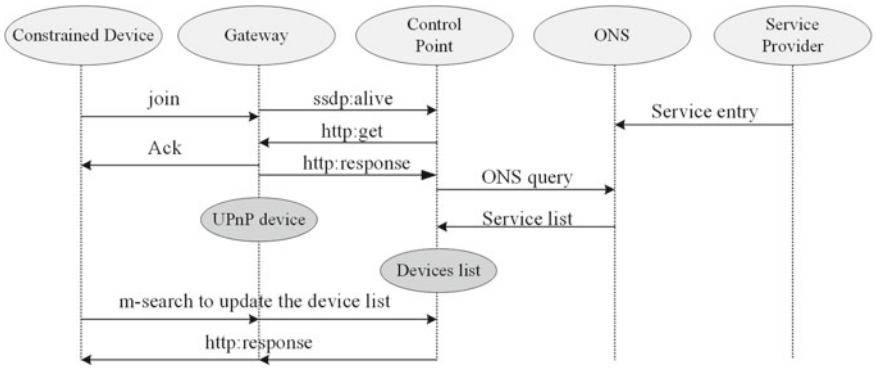


Fig. 19 A devices list is automatically generated in the control point. For constrained devices, a part of capability description may be obtained through ONS

5 Conclusion

In this chapter, an overview of the general PnP IoT architecture and a detailed description of the corresponding components was given. Furthermore, a wide survey was carried out and presented the most important models in the literature that feature PnP capabilities. It is evident that the field has demonstrated significant progress and has already showcased robust and complete PnP solutions. The main industrial players have focused their efforts to develop feature-rich ecosystems that present more and more components with PnP capabilities. However, the ecosystems support vertical architectures restricting the number of applications and IoT solutions that can be developed. In the following years, IoT manufacturers and companies are expected to perform a paradigm shift to interconnected horizontal systems that will benefit significantly the sector. Moreover, future development of PnP architectures will have to tackle challenges with increasing interest such as cybersecurity.

References

1. IEEE Standard for a Simple 32-Bit Backplane Bus: NuBus, ANSI/IEEE Std 1196-1987 (1988)
2. PCI-SIG. PCI Local Bus Specifications. https://pcisig.com/specifications. Accessed 15 Jan 2017
3. MSX Technical Data Book. Hardware/Software Specifications http://map.grauw.nl/resources/system/msxtech.pdf. Accessed 15 Jan 2017
4. IBM's Micro Channel Architecture. http://www.borrett.id.au/computing/art-1989-03-01.htm
5. Fred Krhenbhl: Micro Channel Architecture Bus Master Release 1.1: International Business Machines Corporation (1990)
6. ATmega328/P, 8-bit AVR Microcontrollers complete datasheet, Atmel. http://www.atmel.com/Images/Atmel-42735-8-bit-AVR-Microcontroller-ATmega328-328P_Datasheet.pdf
7. CC3200 SimpleLink Wi-Fi and Internet-of-Things Solution, a Single-Chip Wireless MCU, Texas Instruments. http://www.ti.com/lit/ds/swas032f/swas032f.pdf
8. SAM9G25, SMART ARM-based Embedded MPU datasheet, Atmel. http://www.atmel.com/images/atmel-11032-32-bit-arm926ej-s-microcontroller-sam9g25_datasheet.pdf
9. Raspberry Pi 3 Model B, Raspberry Pi Foundation. https://www.raspberrypi.org/products/raspberry-pi-3-model-b/
10. Adam Osborne, An Introduction to Microcomputers Volume 1: Basic Concepts, Osborne-McGraw Hill Berkeley California USA, 1980 ISBN 0-931988-34-9 pp. 116-126
11. UM10204, I^2C-bus specification and user manual, NXP Semiconductors, Rev.6 - 4 April 2014. http://cache.nxp.com/documents/user_manual/UM10204.pdf
12. How Plug and Play Works. https://technet.microsoft.com/en-us/library/cc781092. Accessed 15 Jan 2017
13. Apache-River Jini Architecture Specification. http://river.apache.org/release-doc/current/specs/html/jini-spec.html. Accessed 15 Jan 2017
14. The Java Virtual Machine Specification, Java SE 7 Edition, Tim Lindholm, Frank Yellin, Gilad Bracha, Alex Buckley, Oracle. Accessed 28 Dec 2013
15. Universal Plug and Play Device Architecture. http://upnp.org/specs/arch/UPnP-arch-DeviceArchitecture-v1.1.pdf. Accessed 15 Jan 2017
16. XML Protocol User Documents, XML Protocol Working Group. https://www.w3.org/2000/xp/Group/
17. User Datagram Protocol: RFC 768, IETF (1980). https://tools.ietf.org/html/rfc768

18. Service Location Protocol, Version 2. https://tools.ietf.org/html/rfc2608. Accessed 15 Jan 2017
19. Open Connectivity Foundation. https://openconnectivity.org/. Accessed 15 Jan 2017
20. IoTivity. https://www.iotivity.org/. Accessed 15 Jan 2017
21. Constrained Application Protocol Specification RFC7252. https://tools.ietf.org/html/rfc7252
22. Shelby, Z., Bormann, C.: 6LoWPAN: The Wireless Embedded Internet, vol. 43. Wiley (2011)
23. Yang, F., Matthys, N., Bachiller, R., Michiels, S., Joosen, W., Hughes, D.: PnP: plug and play peripherals for the internet of things. In: Proceedings of the Tenth European Conference on Computer Systems (EuroSys '15). ACM, New York, NY, USA, Article 25, p. 14 (2015). https://doi.org/10.1145/2741948.2741980
24. Matthys, N et al.: PnP-Mesh: The plug-and-play mesh network for the internet of things. In: IEEE 2nd World Forum on Internet of Things (WF-IoT), Milan, vol. 2015, pp. 311–315 (2015). https://doi.org/10.1109/WF-IoT.2015.7389072
25. Matthys, N., Yang, F., Daniels, W., Joosen, W., Hughes, D.: Demonstration of MicroPnP: the zero-configuration wireless sensing and actuation platform. In: 2016 13th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), London, pp. 1–2 (2016). https://doi.org/10.1109/SAHCN.2016.7732982
26. Yang, F., Ramachandran, G.S., Lawrence, P., Michiels, S., Joosen, W., Hughes, D.: PnP-WAN: Wide area plug and play sensing and actuation with LoRa. In: International SoC Design Conference (ISOCC), Jeju, vol. 2016, pp. 225–226 (2016). https://doi.org/10.1109/ISOCC.2016.7799869
27. Yang, F., Hughes, D., Joosen, W., Man, K.L.: The design of a low-power, high-resolution controller board for PnP. In: International SoC Design Conference (ISOCC), Gyungju, vol. 2015, pp. 173–174 (2015). https://doi.org/10.1109/ISOCC.2015.7401774
28. IEEE Standard for a smart transducer interface for sensors and actuators-common functions, communication protocols, and transducer electronic data sheet (TEDS) formats. In: IEEE Std 1451.0-2007, 21 Sept 2007, pp. 1–335. https://doi.org/10.1109/IEEESTD.2007.4338161
29. IEEE standard for a smart transducer interface for sensors and actuators-network capable application processor (NCAP) information model. In: IEEE Std 1451.1-1999, pp. i (2000) https://doi.org/10.1109/IEEESTD.2000.91313
30. IEEE standard for a smart transducer interface for sensors and actuators-transducer to microprocessor communication protocols and transducer electronic data sheet (TEDS) formats. In: IEEE Std 1451.2-1997, pp. i (1998). https://doi.org/10.1109/IEEESTD.1998.88285
31. IEEE standard for a smart transducer interface for sensors and actuators-digital communication and transducer electronic data sheet (TEDS) formats for distributed multidrop systems. In: IEEE Std 1451.3-2003, 31 Mar 2004, pp. 1–175. https://doi.org/10.1109/IEEESTD.2004.94443
32. IEEE standard for a smart transducer interface for sensors and actuators-mixed-mode communication protocols and transducer electronic data sheet (TEDS) formats. In: IEEE Std 1451.4-2004, pp. 01–430 (2004). https://doi.org/10.1109/IEEESTD.2004.95745
33. IEEE standard for a smart transducer interface for sensors and actuators wireless communication protocols and transducer electronic data sheet (TEDS) formats. In: IEEE Std 1451.5-2007 5 Oct 2007, p. C1-236. https://doi.org/10.1109/IEEESTD.2007.4346346
34. IEEE standard for smart transducer interface for sensors and actuators–transducers to radio frequency identification (RFID) systems communication protocols and transducer electronic data sheet formats. In: IEEE Std 1451.7-2010, 26 June 2010, pp. 1–99. https://doi.org/10.1109/IEEESTD.2010.5494713
35. Nieves, R., Madrid, N.M., Seepold, R., Larrauri, J.M., Arejita Larrinaga, B.: A UPnP service to control and manage IEEE 1451 transducers in control networks. IEEE Trans. Instrum. Meas. **61**(3), 791–800 (2012). https://doi.org/10.1109/TIM.2011.2170501
36. Depari, A., Ferrari, P., Flammini, A., Marioli, D., Taroni, A.: A VHDL model of a IEEE1451.2 smart sensor: characterization and applications. IEEE Sens. J. **7**(5), 619–626 (2007). https://doi.org/10.1109/JSEN.2007.894900
37. Kumar, A., Hancke, G.P.: A Zigbee-based animal health monitoring system. IEEE Sens. J. **15**(1), 610–617 (2015). https://doi.org/10.1109/JSEN.2014.2349073

38. Kumar, A., Hancke, G.P.: An energy-efficient smart comfort sensing system based on the IEEE 1451 standard for green buildings. IEEE Sens. J. **14**(12), 4245–4252 (2014). https://doi.org/10.1109/JSEN.2014.2356651
39. Chi, Q., Yan, H., Zhang, C., Pang, Z., Xu, L.D.: A reconfigurable smart sensor interface for industrial WSN in IoT environment. IEEE Trans. Ind. Inform. **10**(2), 1417–1425 (2014). https://doi.org/10.1109/TII.2014.2306798
40. Becari, W., Ramirez-Fernandez, F.J.: Electrogoniometer sensor with USB connectivity based on the IEEE1451 standard. In: IEEE International Symposium on Consumer Electronics (ISCE), Sao Paulo, vol. 2016, pp. 41–42 (2016). https://doi.org/10.1109/ISCE.2016.7797360
41. Bekan, A., Mohorcic, M., Cinkelj, J., Fortuna, C.: An architecture for fully reconfigurable plug-and-play wireless sensor network testbed. In: IEEE Global Communications Conference (GLOBECOM). San Diego, CA, vol. 2015, pp. 1–7 (2015). https://doi.org/10.1109/GLOCOM.2015.7417564
42. Fortuna, C., Mohorcic, M.: A framework for dynamic composition of communication services. ACM Trans. Sen. Netw. **11**(2), 32:132:43 (2014). https://doi.org/10.1145/2678216
43. Contiki: The Open Source OS for the Internet of Things. http://www.contiki-os.org/
44. RPL: IPv6 Routing Protocol for Low-Power and Lossy Networks, IETF. https://tools.ietf.org/html/rfc6550. Accessed Mar 2012
45. Mikhaylov, K., Huttunen, M.: Modular wireless sensor and Actuator Network Nodes with Plug-and-Play module connection. In: IEEE SENSORS: Proceedings Valencia **2014**, 470–473 (2014). https://doi.org/10.1109/ICSENS.2014.6985037
46. Mikhaylov, K., Paatelma, A.: Enabling modular plug-n-play wireless sensor and actuator network nodes: software architecture. In: IEEE SENSORS, Busan, vol. 2015, pp. 1–4 (2015). https://doi.org/10.1109/ICSENS.2015.7370252
47. Yang, F., Hughes, D., Matthys, N., Man, K.L.: The PnP Web Tag: A plug-and-play programming model for connecting IoT devices to the web of things. In: 2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), Jeju, South Korea, pp. 452–455 (2016). https://doi.org/10.1109/APCCAS.2016.7804000
48. Bordel, B., Rivera, D.S.D., Alcarria, R.: Plug-and-play transducers in cyber-physical systems for device-driven applications. In: 2016 10th International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing (IMIS), Fukuoka, pp. 316–321 (2016). https://doi.org/10.1109/IMIS.2016.68
49. Dai, W., Huang, W., Vyatkin, V.: Enabling plug-and-play software components in industrial cyber-physical systems by adopting service-oriented architecture paradigm. In: IECON 2016-42nd Annual Conference of the IEEE Industrial Electronics Society, Florence, pp. 5253–5258 (2016). https://doi.org/10.1109/IECON.2016.7793834
50. IEC 61499: Function Blocks, International Standard, 2nd edn (2012)
51. Erl, T.: Service-Oriented Architecture: Concepts, Technology and Design, 760 pp. Prentice Hall Professional Technical Reference (2005)
52. Chan, S., Kaler, C., Kuehnel, T., Regnier, A., Roe, B. Sather, D., Schlimmer, J.: Devices Profile for Web Services. Microsoft Developers Network Library, Feb 2006
53. Khan, M.S., Islam, M.S., Deng, H.: Design of a reconfigurable rfid sensing tag as a generic sensing platform toward the future internet of things. IEEE Internet Things J. **1**(4), 300–310 (2014). https://doi.org/10.1109/JIOT.2014.2329189
54. EPC Radio-Frequency Identity Protocols Class-1 Generation-2 UHF RFID Conformance Requirements Version 1.0.5: GS1 EPCglobal. http://www.gs1.org/sites/default/files/docs/epc/uhfc1g2_1_1_0-Conformance%20Test%20Methods_1_0_5-20070323.pdf. Accessed Mar 2007
55. EPCTM 9 Generation 1 Tag Data Standards Version 1.1 Rev.1.27: GS1 EPC Global. http://www.gs1.org/sites/default/files/docs/epc/tds_1_1_rev_1_27-standard-20050510.pdf. Accessed May 2005
56. Cirani, S., et al.: A scalable and self-configuring architecture for service discovery in the internet of things. IEEE Internet Things J. **1**(5), 508–521 (2014). https://doi.org/10.1109/JIOT.2014.2358296

57. Tseng, C.W., Chen, Y.C., Huang, C.H.: Integrating transducer capability to GS1 EPCglobal identify layer for IoT applications. IEEE Sens. J. **15**(10), 5404–5415 (2015). https://doi.org/10.1109/JSEN.2015.2438074
58. Perera, C., Jayaraman, P.P., Zaslavsky, A., Georgakopoulos, D., Christen, P.: Sensor discovery and configuration framework for the internet of things paradigm. In: IEEE World Forum on Internet of Things (WF-IoT), Seoul vol. 2014, pp. 94–99 (2014). https://doi.org/10.1109/WF-IoT.2014.6803127
59. Houyou, A.M., Huth, H.P., Kloukinas, C., Trsek, H., Rotondi, D.: Agile manufacturing: general challenges and an IoTWork perspective. In: Proceedings of 2012 IEEE 17th International Conference on Emerging Technologies & Factory Automation (ETFA 2012), Krakow, pp. 1–7 (2012). https://doi.org/10.1109/ETFA.2012.6489653
60. Dai, L., Liu, L., Sun, X.: The system construction and research review of agile manufacturing. In: 2011 International Conference on Management and Service Science, Wuhan, pp. 1–4 (2011). https://doi.org/10.1109/ICMSS.2011.5998319
61. Rotondi, D., et. al.: D1.1 State of the art and functional requirements in manufacturing and automation; IoT@Work public deliverable. https://www.iot-at-work.eu/downloads.html. Accessed 30 Dec 2010
62. Houyou, A., Huth, H.-P.: Internet of things at work European project: enabling plug&work in automation networks. In: Embedded World Conference 2011, Nuremberg, Germany (2011)
63. Mitsugi, J., Sato, Y., Ozawa, M., Suzuki, S.: An integrated device and service discovery with UPnP and ONS to facilitate the composition of smart home applications. In: IEEE World Forum on Internet of Things (WF-IoT), Seoul vol. 2014, pp. 400–404 (2014). https://doi.org/10.1109/WF-IoT.2014.6803199
64. EPCglobal Object Name Service (ONS) 1.0.1 (2008)
65. Mitsugi, J., Yonemura, S., Yokoishi, T.: Reliable and swift device discovery in consolidated IP and ZigBee home networks. IEICE Trans B **E96-B**(07) (2013)

Digital Forensics for IoT and WSNs

Umit Karabiyik and Kemal Akkaya

Abstract In the last decade, wireless sensor networks (WSNs) and Internet-of-Things (IoT) devices are proliferated in many domains including critical infrastructures such as energy, transportation and manufacturing. Consequently, most of the daily operations now rely on the data coming from wireless sensors or IoT devices and their actions. In addition, personal IoT devices are heavily used for social media applications, which connect people as well as all critical infrastructures to each other under the cyber domain. However, this connectedness also comes with the risk of increasing number of cyber attacks through WSNs and/or IoT. While a significant research has been dedicated to secure WSN/IoT, this still indicates that there needs to be forensics mechanisms to be able to conduct investigations and analysis. In particular, understanding what has happened after a failure or an attack is crucial to many businesses, which rely on WSN/IoT applications. Therefore, there is a great interest and need for understanding digital forensics applications in WSN and IoT realms. This chapter fills this gap by providing an overview and classification of digital forensics research and applications in these emerging domains in a comprehensive manner. In addition to analyzing the technical challenges, the chapter provides a survey of the existing efforts from the device level to network level while also pointing out future research opportunities.

U. Karabiyik (✉)
Department of Computer and Information Technology, Purdue University, Knoy Hall, Room 225 401 N. Grant St., West Lafayette, IN 47907, USA
e-mail: ukarabiy@purdue.edu

K. Akkaya
Department of Electrical and Computer Engineering, Florida International University, Miami, FL 33174, USA
e-mail: kakkaya@fiu.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_6

1 Introduction

Wireless Sensor Networks (WSNs) have been initially proposed for military operations by the end of 90s [27]. However, with their potential in many applications, they have started to be deployed in different civil applications in early 2000s. WSNs have been touted to be used in many applications. These include but is not limited to environmental monitoring, habitat monitoring, structural health monitoring, health applications, agriculture applications, and surveillance [93]. Typically, in such applications, a large number of sensors are deployed to sense the environment and send the collected data to a gateway or base-station for further processing. The communication is multi-hop and all the nodes are assumed to be battery operated with limited processing and storage capabilities. There has always been incredible interest in WSN research from node level to application level [3]. The bulk of WSN research has focused on energy-efficient protocols at different layers of the protocol stack. The goal was to maximize the lifetime of the WSNs while enabling distributed operations. Energy-efficient MAC, routing, and transport protocols have been proposed [2, 22, 89]. Later, these protocols were augmented with security capabilities [88]. Despite the huge amount of research, the WSN market was not mature. In the early 2000s, there were only a few sensor products (such as Mica2) and standardization efforts were not adequate. Therefore, the use of term WSN has been diminished and efforts are directed toward more personal sensor devices that came with the proliferation of smartphones and other wearable devices. The attention has been shifted to these devices, referred to as Internet of Things (IoT).

The term IoT was first phrased in the context of supply chain management by Kevin Ashton in 1999 to get executive attention at Procter and Gamble [8]. Although it was used in different and somewhat related concepts earlier, the definition has become more comprehensive to comprise devices from health care to entertainment and transportation to building management [80]. Therefore, the term might be used to describe the world where other devices are uniquely distinguishable, addressable, and contactable by means of the Internet. For example, smart homes are furnished with hi-tech devices controlling such devices as the TV, refrigerator, microwave, blinds, music system, air conditioning units.

Today, we have more than 5 billion "things" connected to the Internet and this number is expected to be nearly 50 billion (there are also different estimates) by 2020 [86]. Taking the advantage of using RFID and sensor network technology, physical objects such as computers, phones, wearable technologies, home appliances, vehicles, medical devices, and industrial systems can be easily connected, tracked, and managed by a single system Jiang et al. [37]. One of the many reasons to get these devices connected is that most of the people want to take advantage of being conveniently "online" in this age of Internet. On the other hand, we do underestimate the downsides of being connected in every second of every day.

Although the expected number of connected devices is hypothetical, there is a real issue regarding the existence of such a large collection of devices, which are mostly vulnerable to cyberattacks. On October 21, 2016, we faced the reality of how our

innocent household devices connected to the Internet could be part of an IoT army committing Distributed Denial of Service attack (DDoS) to shut down websites including Twitter, Netflix, PayPal, and Amazon Web services [91]. In addition to being vulnerable, Syed Zaeem Hosain, CTO of Aeris—a pioneer in the machine-to-machine market, has raised the concern that scalability in IoT is the biggest issue as such a large number of devices will be generating enormously big data [33]. Hence, the following questions are asked by Mr. Hosain:

- How will we transport such large data?
- How will we store it?
- How will we analyze it?
- How will we search/find targeted data in large collection?
- How will we keep the data secure and private?

All of these questions are part of our concerns about IoT today, however it is urgent that these issues must be addressed in advance before we are faced with serious scalability issues.

Miorandi et al. have discussed that IoT is a leading technology, which brings various areas from cyber and the physical world together by means of making physical devices smarter and connected with one another [53]. By taking this into account, the usage of the term IoT can be generalized into the following broad areas as discussed in Atzori et al. [9], Peña-López [64]:

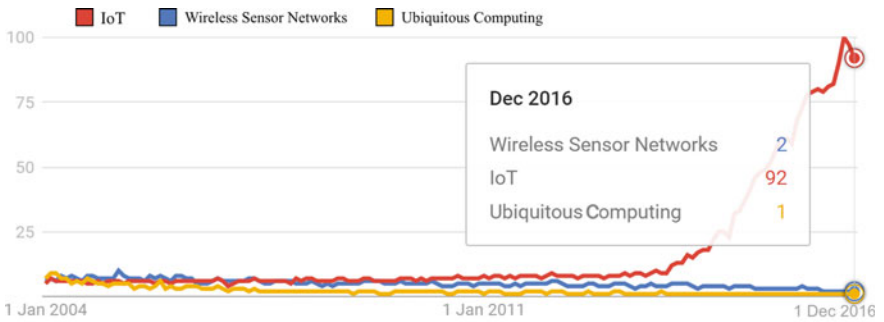
- the global network providing an ultimate interconnection ability to the smart things (devices) via the Internet
- The collection of assistive technologies (e.g., RFIDs, Near-Field Communication devices, and WISP.)
- The group of applications and services (e.g., Cloud services and Web of things.)

Although different terms have similar meaning the popularity of concepts has changed over the time. Now we look at web search popularity of the terms IoT (including Internet of Things), WSNs and Ubiquitous Computing (UC) as they are used interchangeably. Figure 1a is created using Google Trends and it shows how fast the IoT popularity has increased in web searches compared to the terms WSNs and UC over the decade. Similarly, Fig. 1b shows how the popularity of the terms WSNs and UC has decreased (comparatively) since 2004.

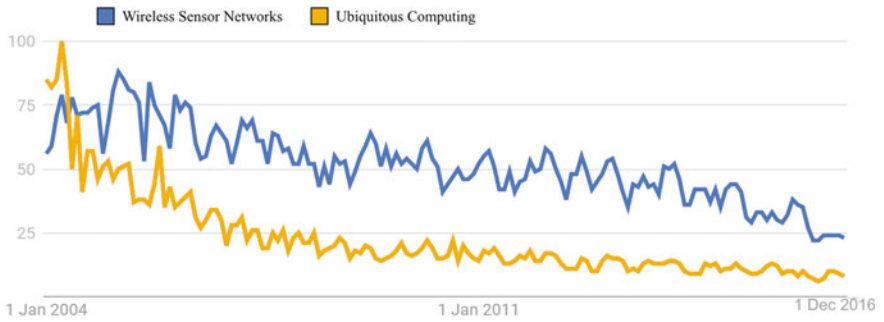
Whether it is called IoT or WSN, there has been a lot of studies to secure these networks starting from the node level to network level [96]. The security services provided in IoT and WSN include confidentiality, integrity, authentication, access control, anonymity, and availability.

However, with the increasing prevalence of these devices in many real-life applications, a need has emerged for conducting digital/network forensics to be able to understand the reasons for failures and various attacks. Therefore, in recent years also we have witnessed some studies on cyber forensics that relate to WSNs or IoT. The goal of this chapter is to investigate such forensic research on WSNs and IoT, and put them in a systematic manner for better understanding and future research.

This chapter is organized as follows: In the Sect. 2, we provide a brief background on digital forensics. Section 3 presents related background in IoT and WSNs.



(a) Search trends for IoT, Wireless Sensor Networks and Ubiquitous Computing



(b) Search trends for Wireless Sensor Networks and Ubiquitous Computing

Fig. 1 Google search trends between January 2004 and December 2016: The numbers represent search interest relative to the highest point on the chart for all world and time. A value of 100 is the peak popularity for the term. A value of 50 means that the term is half as popular. Likewise, a score of 0 means the term was less than 1% as popular as the peak [30]

Section 4 discusses how digital forensics can be applied to IoT and WSN environments.

2 Digital Forensics

Digital Forensics is a branch of forensics science particularly targeting identification, collection (a.k.a. acquisition), examination, analysis, and reporting of digital evidence in order to present it to a court of law. Figure 2 shows the U.S. Department of Justice's digital forensics investigative process described in "A guide to first responders" [24]. Digital forensics investigators deal with tremendous amounts of data from numerous types of devices including computers, phones, wearable devices, industrial controls systems, military deployment systems.

Preserving the integrity of evidence in an essential duty in order to make sure the collected evidence is forensically sound. When a crime/incident occurs, incident first

responders arrive to the scene to identify and secure the digital devices to preserve the forensics soundness of the evidence. After securing the evidence devices digital forensics investigators collect digital evidence for further examination and analysis. This basically means to find crime/incident-related data on the digital device such as finding traces of an attack and its timestamp on memory of hacked smart TV. During the collection, examination, and analysis phases, investigators use digital forensics tools (both hardware and software). These tools help investigators to locate and recover digital evidence which can be both inculpatory (evidence that proves the guilt) and exculpatory (evidence that proves the innocence). At the reporting phase, investigators prepare a report to include in their testimony. When the investigator is asked to testify and present the evidence at a court, the admissibility of the evidence will be questioned based on the procedures followed by the investigator. The most important factor for the admissibility is to verify that the evidence device has not been altered during the investigation. In the case of the IoT environment, this may be quite challenging as there is no universal standard to collect, examine and analyze data from IoT.

Due to the accelerated advancement in technology, particularly in the past two decades, huge numbers of (heterogeneous) objects became available for personal or enterprise use. This also yields an enormous amount of heterogeneous data and thus more sophisticated and more difficult digital forensics investigations.

3 Related Background on IoT and WSNs

The evolutionary background of IoT lies in the advancement of the technology on microsensor devices in the later 90s. Specifically, the advancements in microprocessors, memory technology, and more importantly micro-sensing devices led to the development of tiny sensors. These sensors are then equipped with radio communication capability on battery energy, which enabled unattended intelligent sensing devices that can gather, process, and transmit data. In the early 2000s, many sensor devices were built to fit the needs of various applications as seen in Fig. 3.

Of particular interest to these devices are their resources, especially in terms of memory and storage. Early sensors have very scarce resources in terms of memory, which makes data storage almost impossible for forensics purposes. Typically,

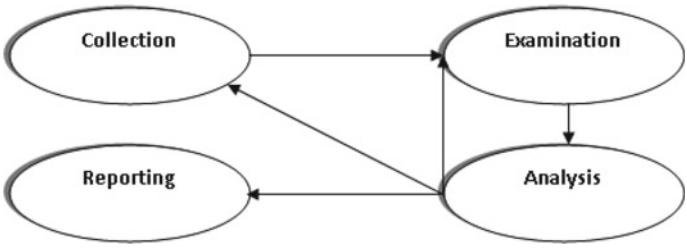


Fig. 2 Digital forensics process model [39]

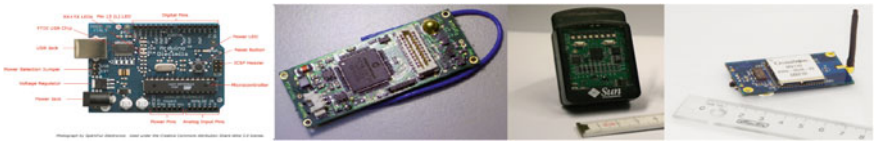


Fig. 3 Various sensor platforms: Arduino, Particle, SUN Spots, and IRIS Motes

there were two cases where memory was used: (1) User memory used for storing application-related or personal data; and (2) Program memory used for programming the device. This memory also contains identification data if the device has any. Table 1 shows the memory sizes for these devices.

The other sensor component that is of interest due to forensics would be the communication module. Early devices relied on energy-aware MAC protocols for communication [22]. Some of them later became standards such as Zigbee [12], but some of them were only adapted such as Bluetooth [13] as seen in Table 2.

At the network layer, energy-efficient routing protocols were developed to provide service to large-scale WSNs consisting of thousands of nodes [6, 83, 90] and employ multi-hop communication [2]. ZigBee Alliance had also routing and application layer protocols for WSNs [12]. In most cases, these protocols were distributed and required sensor nodes to maintain a simple routing table for data forwarding. In some cases, the routing protocol was managed by the gateway which is assumed to collect all the sensor data from the sensors. In any case, there was enough information in the sensors or gateway to be able to identify routing failures in real time, but this might be challenging for cyber forensics purposes as will be discussed later.

The heavy research on WSNs later led to the development of some standards such as ZigBee/IEEE802.15.4, IETF ROLL [42, 73], IETF 6LoWPAN [72], Wireless

Table 1 Memory size for different platforms

Devices	Memory size
Passive tags	O (100 B)
Active tags	O (1 kB)
IRIS motes	O (100 kB)
Gateways	O (100 MB)

Table 2 MAC protocols and data rates for different sensors

Standard	Rate	Frequency
Bluetooth	2.1 Mb/s	2.4 GHz
ZigBee	250 kb/s	2.4 GHz/918 MHz/868 MHz
RFM TR 1000 (proprietary)	19.2 kb/s	916.5 MHz
Chipcon CC 1000 (proprietary)	100 kb/s	433 MHz

HART [77], ISA100 [19] which accelerated the production of sensor devices. In the meantime, there has been further developments to enrich the resources of sensor devices and getting them connected to the Internet. The enrichment was in terms of processor and memory capacity and the number of sensing modules. With the proliferation of smart mobile phones, the idea of smart, connected, sensing, and battery-operated devices have penetrated our everyday lives which has led to the development of similar products to make our lives convenient. Within a few years, a lot of sensing and communication capable devices such as smart meters, cameras, thermostats, wearables, RFIDs, tags, bulbs, beds, speakers, locks, watches, cookers, keypad, and appliances, have started to be seen which are referred to as IoT devices in general [23].

With the enriched resources, these devices started to look like more of our laptops with comparable memory/storage sizes and communication capabilities. In addition to ZigBee or Bluetooth, WiFi/4G has also been started to be used for communication purposes. Finally, the data collected from these devices was not stored in the gateways but rather transferred to cloud storage where it can be accessed for later use.

The IoT era changed the needs of the WSN era and Digital Forensics was one of the affected area as the devices are being used in a lot of daily applications by humans. Therefore, we discuss how digital forensics is applied to the IoT and WSNs.

4 Applying Digital Forensics to IoT and WSNs

IoT and WSNs consist of sensitive data stored and processed hence, in theory, it is suggested that the data which is processed and cumulated by well-known firms will be the subject of future digital forensics investigations. The evidence that is provided by IoT or WSNs to the forensics community will be far more finer compared to what the community currently possesses. In addition, IoT and WSNs also offer new and better opportunities for data that is at times misused, through growth and development in the forensics community's procedures. The techniques/algorithms methods that were used and or developed were based on the digital forensics process model consisting of collection, examination, analysis, and reporting of the data/evidence. Using these practices not only data for evidence is identified in a myriad system, but is also preserved for future references as the information presented is an intense fusion of collection, extraction, processing, and interpretations.

Digital forensics in IoT/WSNs is a challenge, especially when it comes to accuracy due to the intensity of analysis. This results in data sometimes losing its granularity as systems may store, use, or present different semantics, however, it does have the ability to adopt dissimilar formats, and may hold a proprietary format. Taking into the heterogeneity of data that IoT/WSNs devices generate it is even more challenging. The following questions must be answered before the investigation is being performed in order to avoid inadmissibility of evidence. Can data be collected from the devices using available tools? Is the data propriety? How can it be analyzed? Are forensic tools compatible with this data?

Most of the challenges in IoT forensics are also available to the WSNs particularly at the device/data storage and network levels. The only difference in most cases is the scale of WSNs because of the application-specific needs. Early WSNs works lacked any security in regards to integrity and authentication because of the broadcast nature of communication. There was no formal set of requirements for achieving forensic readiness in WSNs. However, with the rapid tendency toward the usage of efficient, low memory footprint, and low-power devices in the industry, devices will be less likely to keep data stored in memory. Therefore, similar forensic readiness frameworks that will be discussed in the following sections must be developed for such devices in advance. Otherwise, forensically crucial data can be easily lost forever.

In the next section, we will discuss the challenges that investigators and practitioners face when performing digital forensics procedures in both IoT and WSNs. Although IoT and WSNs are different with respect to their structures, WSNs are considered to be part of IoT [17, 38, 48, 50], a concept of worldwide connected ubiquitous devices. The distinctions between two environments are not clearly pointed by the research in the current literature and digital forensics efforts similarly applied to both concepts, particularly research in IoT forensics are conducted with WSN characteristics in mind. This makes some of the research efforts in both the environments inseparable from the digital forensics perspectives.

4.1 Challenges in IoT and WSN Forensics

In this section, we discuss the digital forensics challenges for IoT and WSN as specified by Hegarty et al. [32]. Note that most of the challenges we discuss in this section are applicable to both IoT and WSN.

Different Interfaces and Storage Units: The IoT devices that are used in everyday life have different interfaces, which allow users to use services or control the devices. Example interfaces could be propriety software, mobile application, hardware, or embedded firmware which provides an invisible interface. The variety of interfaces makes digital forensics investigation a tedious process as digital forensics tools do not automatically detect all types of interfaces, file systems, and even data itself. Similar issues arise when WSNs are the forensically targeted environments. In addition to the variety of interfaces, IoT devices store data in miscellany of storage units both volatile and nonvolatile including internal and external memory units (e.g., eMMC, eFlash, and DRAM) and cloud storage (e.g., HDD and SSD) [65]. As for the sources of digital evidence, Table 3 gives a broader view of where potential evidence may reside in an IoT and WSN environment.

Differences in the interfaces and storage units cause investigators to perform manual forensic methods on the devices if (at all) possible. This will also increase the time required for the investigation as automated tools do not recognize propriety interfaces. Another issue is that volatile data might be destroyed by the device after they are used. In this case, data recovery may not be even possible. In addition,

Table 3 Potential evidence sources in IoT and WSN environments [62]

	Sources	Example	Expected evidence
Internal to network	Hardware end nodes	IoTware, e.g., game consoles, fridges, mobile devices, smart meters, readers, tags, embedded systems, heat controller	Sensor data, e.g., IP address, Rime number, sensor ID
	Network	Wired and Wireless, mobile communications, e.g., GSM, sensor networks, HIDS, NIDS, HMS	Network, logs
	Perimeter devices	AAA server, firewall, NAT server, IDS, NIDS, HIDS	
External	Cloud	Public, private, hybrid cloud systems	Client virtual machines; logs
	Web	Web clients, webservers, social networks	Web logs; user
	Hardware end nodes	Mobile devices, sensor nodes and networks	Sensor data e.g. IP address, rime number, sensor ID
	'X' Area Networks	Home area networks (HAN)	Network logs

data may also be destroyed due to the wear-leveling technology in flash memory devices and solid-state drives. Every memory cell has a certain read/write lifetime which varies between 10,000 and 100,000 depending on the manufacturer. Therefore, internal firmware in the memory will distribute data stored in the memory to the unused (unallocated) cells in order to level memory wearing in mostly used memory cells. In this case, previously deleted data will be destroyed because unallocated space also consists of the memory cells that has previously been used to store data but later deleted.

Furthermore, data stored in the cloud raises serious issues in digital forensics investigations performed in IoT and WSN environment. In order to identify these issues, [69] defines cloud forensics in three dimensions (see Fig. 4). The two most important problems in cloud forensics that are also highly related to IoT and WSN forensics are multi-tenancy and multi-jurisdiction. Multi-tenancy allows cloud tenants to access the software instance simultaneously [20], therefore, user ascription and ownership for specific data become the investigators' major concerns.

In order to provide efficient service availability and reduce the cost of services, major cloud service providers such as Google, Amazon, and HP locate their data

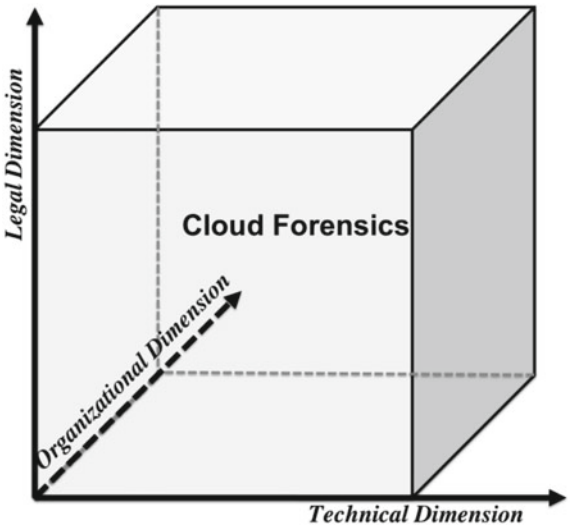


Fig. 4 Cloud forensic three-dimensional model [69]

centers all around the world. Different countries and different states have different jurisdictions. A crime in one jurisdiction may not be considered a crime in another. In addition, law enforcement agencies having different jurisdictions may not be willing to cooperate with each other. Due to all of these issues, investigators may have to deal with multi-jurisdiction issues when data from IoT and WSNs are stored in cloud.

Lack of Universal Standard for IoT and WSN Data Storage Due to the characteristics of IoT/WSN data, it is extremely difficult to create a universal standard for data storage. Nevertheless, there have been efforts to create frameworks to provide a unified way to store data for IoT. Li et al. [49] identifies the IoT data features as follows:

- ***Multi-source and Heterogeneity***: IoT/WSN data is sampled by various connected devices including Radio-Frequency Identification (RFID) readers, cameras, smart appliances, proximity, pressure, temperature, humidity, and smoke sensors. The data collected from this vast category of devices have significantly different semantics and structures.
- ***Huge scale***: The IoT/WSN contains a large number of perception devices, these devices' continuously and automatically collect information leads to a rapid expansion of data scale.
- ***Temporal–spatial correlation***: As the data are constantly collected from IoT/WSN, the data will consist both time and space attributes in order to correlate them with respect to the changing location of device over time.
- ***Interoperability***: IoT/WSN are currently evolving to achieve data sharing to facilitate collaborative work between different applications. For instance, in the case of an on-road emergency, while the patient's medical record is securely shared with a nearest emergency center [67], the data related to road conditions may be also assessed for timely arrival by an autonomous car.

- ***Multidimensional***: IoT application now integrates several sensors or WNSs to simultaneously monitor a number of sensing devices, such as temperature, humidity, light, pressure, and so on. And, thus the sample data is usually multidimensional.

The available methods and techniques are mostly limited and designed for a certain set of technologies. For instance, Li et al. [49] have proposed a solution to the storage and management of IoT data named IOTMDB using NoSQL (Not Only SQL). In addition to this work, Jiang et al. [37] proposed a data storage framework to efficiently store big IoT data which is collected from the deployed devices (WSNs) into storage units by combining and extending multiple databases and *Hadoop* (an open-source framework that provides capability of process and storage of large data sets). In addition, Gubbi et al. [31] introduced a conceptual IoT framework with *Aneka* cloud computing platform–runtime platform and a framework for developing distributed applications on the cloud [51]—being at the center. This framework integrates ubiquitous sensors and various applications (e.g., surveillance, health monitoring, and critical infrastructure monitoring) using aforementioned cloud platform.

From the forensics investigation's point of view, analysis of data coming from different sources will be a serious challenge. The only way to deal with the analysis of such heterogeneous data is to use Hexadecimal editors (a.k.a HEX editors) as they allow reading the raw data from storage units. However, it will be a tedious (if not infeasible with large-scale data) process because of the amount of data collected from IoT devices and WSNs.

Temporal–spatial correlation of IoT/WSN data may be useful for the investigators when data includes geolocation information (e.g., GPS coordinates) readable by the tools used. However, IoT/WSN space can be defined of any size and data may come with custom space information. This also needs to be translated into intelligible data by the investigators as evidence.

Interoperability of the devices will be a serious challenge for forensics investigations as the data will be shared among the applications and the origin of the data needs to be known to conclude the investigation. If the data being operated by different applications is not traceable then accountability or non-repudiation issues will be raised. For instance, it will be difficult to answer the questions: What caused the operation failure? Was there any attack? What data is produced by each application/device?

Devices have different levels of complexity, battery life/source As discussed earlier, IoT devices may vary depending on the duties they perform, how often the device communicates, size of the data being transmitted, and available storage in the device [75]. This variance is also reflected in the complexity of devices. While the device may be as simple as a single sensor collecting environmental values from animals' habitats, it may also be complex enough to consist of a processor, relatively large memory units, and communication protocols with security mechanisms (e.g., Internet refrigerator). In the former, battery replacement will be impractical, therefore battery life is expected to outlive the animal [16]. In the latter however, the device will need to constantly consume power to be available for its service.

Table 4 Comparison tables for different IoT and WSN operating systems and supported protocols [52]

OS	Min RAM	Min ROM	C Support	C++ Support	Multi-Threading	Modularity	Real-Time
Contiki	<2KB	<30KB	P	N	P	P	P
TinyOS	<1kB	<4kB	N	N	P	N	N
RIOT	∼1.5kB	∼5kB	Y	Y	Y	Y	Y
Linux	∼1MB	∼1MB	Y	Y	Y	Y	P

OS	IPv6	TCP	6LoWPAN	RPL	CoAP
Contiki	Y	P	Y	Y	Y
TinyOS	N	P	Y	Y	Y
RIOT	Y	Y	Y	P	P
Linux	Y	Y	Y	Y	N

Complexity and battery life/source of the device affect digital forensics investigations from similar points of view as discussed above such as volatility of data, availability of data, ownership, and user ascription. For example, the data in network and volatile memory disappears in a short amount of time, thus recovery of such data is often impossible unless the device keeps logs of data. This requires existence of more nonvolatile memory and processing power hence larger battery.

Availability of Propriety Operating Systems The operating systems (OS) that are used for IoT was originally designed for WSNs such as TinyOS [47], Contiki [25] and OpenEmbedded Linux [61]. However, with the advances in the development of more sophisticated IoT devices than small sensors, the need for new OSs for IoT emerged. Hence, RIOT [11] was developed to bridge the gap between the available OSs for WSNs and the new needs for IoT. Recent development of Android Things [29] also move this trend to another level to provide leveraging the existing Android development tools, APIs and resources to build an IoT environment. While Table 4 depicts the comparison of different OSs for IoT and WSN, the details about these existing solutions can be obtained from the given resources above. In Table 4, P means: Partial Support, N means: No Support and Y means: Full Support for given points.

There has been digital forensics research on the protocols such as IPv6 [45, 46, 60], 6LoWPAN [46, 66], and RPL [46] that we mentioned in Table 4. These research efforts mostly provide frameworks for forensic readiness of IoT and WSNs. To the contrary of the availability of forensic readiness frameworks, wide variety of available protocols for both IoT and WSN creates a troublesome investigative process and introduces a steep learning curve for forensic examiners.

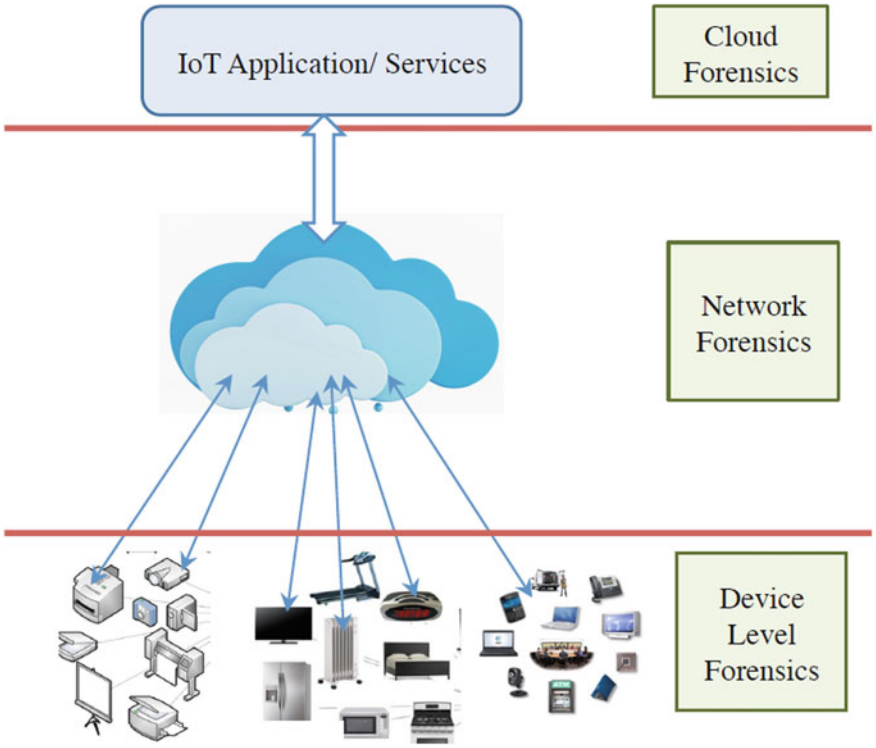


Fig. 5 IoT forensics [96]

IoT forensics can be divided into three categories depending on where the forensic data is located and the investigation can take place (see Fig. 5) [96]. Specifically, these are : (1) Device/Node; (2) Network where the data is collected; and (3) Cloud where the data is stored. The forensic research on WSNs is conducted at the first and second levels where sensor data is collected and transferred, and the communication takes place. Next we explain each category.

4.2 Device-Level Investigation

IoT or WSNs deploy a variety of devices with certain characteristics. Typically these devices employ processing units, memory, communication module, and sensing modules. The richness of the set of such devices increased significantly with the developments in micro-electromechanical devices [28]. Examples of devices include but is not limited to sensors, smartphones, smart meters, smart thermostats, cameras, wearable devices, on-board vehicle devices, RFIDs, smart watches, and drones.

Device-level investigation is necessary when data needs to be collected from the memory of a device in IoT/WSN. As discussed in Sect. 4.1, IoT/WSN devices may have propriety interfaces and storage units. Although it creates a burden on investigators in terms of longer investigation time and increased learning curve, evidence must be collected from these heterogeneous devices. Thus, the current state of the research shows that there is a long way to standardize the device-level investigations for both IoT and WSNs environments. In this section, we explain general forensics techniques, which are used when data is not available through device's interface. We then discuss some of the techniques used to collect forensic data from specific devices and their memories.

The National Institute of Standards and Technology (NIST) discussed different digital forensics data acquisition techniques from mobile devices in "Guidelines on Mobile Device Forensics" [10]. They recommend performing the following acquisition methods: manual extraction, logical extraction, hex dumping/JTAG, chip-off, micro read. Manual and logical acquisition methods are usually available when devices provide user interface and are not locked, password protected, and damaged. In the case of IoT devices (other than smartphones and tablets), it is mostly not the case. Therefore, investigators usually perform hex dumping/JTAG and chip-off techniques (micro read is a special technique and it is out of our scope).

When smartphones or tablets are the interests of the investigations, examiners use state-of-art digital forensic tools such as Cellebrite UFED Physical Analyzer, Paraben Device Seizure, XRY, and Oxygen Forensics for their data acquisition and analysis. This is mainly because these devices come with a well understood operating systems such as Android, iOS, or Windows. Therefore, physical and logical acquisition is generally available to the investigators using aforementioned toolkits. Although we discuss some data acquisition techniques from mobile devices in this section, we do not elaborate more on the available forensic toolkits.

Forensically related data from a mobile device's main storage unit is typically available for acquisition, however volatile data acquisition could often be challenging. Therefore, particular research interest from the mobile forensics community emerged for volatile memory acquisition. Anderson [5], Kollár [43], Sylve et al. [82] proposed early forensic volatile memory dumping tools *crash*, *fmem*, and *dmd*, respectively. The acquired data from these tools is then analyzed using other available tools such as hex editors. As a more recent research, Saltaformaggio et al. [70] proposed an open-source tool called RetroScope, which recovers multiple previous screens (from 3 to 11) from the volatile memory of a smartphone using a spatial–temporal memory acquisition technique. This technique shows that investigators can recover earlier content of an application (e.g., Facebook, WhatsApp, and WeChat) after the data is not available through conventional techniques and tools. This technique can also be particularly effective when the investigators do not have access to the smartphone's data storage due to being password protected. Another recent research on volatile memory acquisition tool development for mobile devices is done by Yang et al. [94]. The proposed tool, AMExtractor, collects volatile memory from a wide variety of Android devices for forensic acquisition meaning with high integrity.

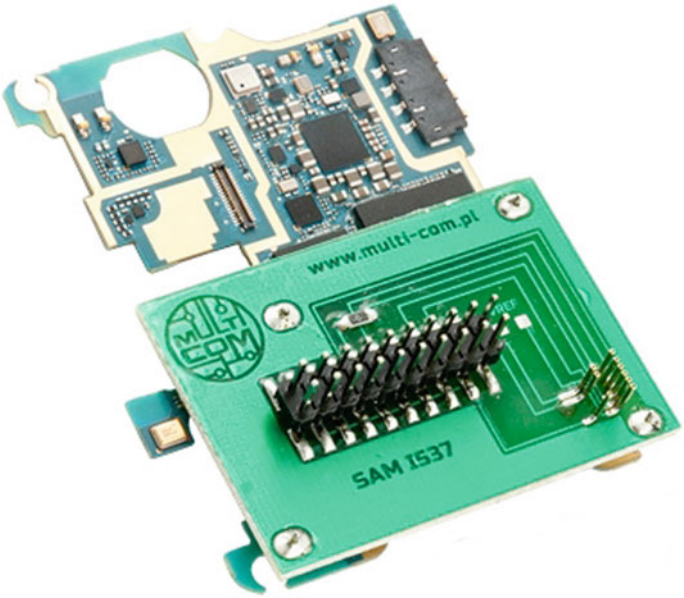
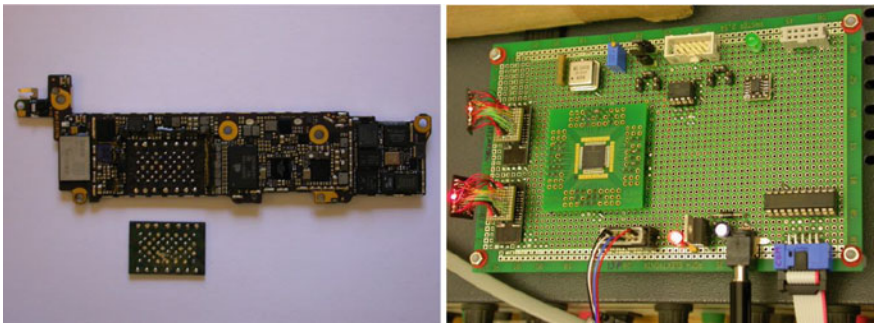


Fig. 6 JTAG module for Samsung Galaxy S4 active smartphone

Hex dumping/JTAG technique allows investigators to access the memory content by connecting special cables to the provided pins on the device. This is done by loading a firmware to the device's memory which is then used to access the information in the rest of the device memory. Figure 6 shows the JTAG module attached to Samsung Galaxy S4 Active phone's mainboard. Using the connectors available on the module and forensic memory reading tools, data from the phone's memory can be easily accessed.

Chip-off is another technique used when phone data is not available due to several reasons such as JTAG is not possible and phones being physically broken, burned, or locked. In such cases, investigators can physically remove the flash memory from the device using chip-off technique. Although this technique is described for mobile phones, it can also be used pretty much for any IoT device or a sensor in WSNs which stores data in flash memory (NAND, NOR, OneNAND, or eMMC) [35]. It is also important to note that chip-off techniques may damage the memory and may cause permanent data loss even though all the precautions are taken [81].

Chip-off is a delicate and challenging method of data acquisition, therefore, it requires extensive training in both electronic engineering and file system forensics. After the memory is removed from the phone, investigators are able to create binary image (bit-by-bit copy) of the removed memory. Figure 7a shows removed NAND flash memory from iPhone 5c and Fig. 7b shows an example of how removed NAND chip is mirrored using a test board. Finally, Fig. 8 shows the raw data acquired from the removed memory via chip programmer and reading program.



(a) iPhone 5c with removed NAND (b) Test board for copying NAND chips

Fig. 7 Chip-off and NAND memory mirroring for iPhone 5c (both figures are from [76])

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F	ASCII
00	6C	6C	6F	63	61	74	6F	72	49	53	33	5F	45	45	45	44	llocatorIS3_EEED
10	31	45	76	00	5F	5F	5A	4E	53	74	33	5F	5F	31	32	30	1Ev __ZNSt3__120
20	5F	5F	73	68	61	72	65	64	5F	70	74	72	5F	70	6F	69	__shared_ptr_poi
30	01	01	8A	AE	03	09	01	00	1E	5A	00	00	03	09	01	00	☺☺Ko♥☐☺ ▲Z ♥☐☺
40	6E	74	65	72	49	50	4E	33	70	63	69	39	69	6E	74	65	nterIPN3pci9inte
50	72	66	61	63	65	37	63	6F	6E	74	72	6F	6C	31	32	63	rface7control2c
60	6C	69	65	6E	74	42	75	66	66	65	72	45	4E	53	5F	31	lientBufferENS_1
70	34	64	65	66	61	75	6C	74	5F	64	65	6C	65	74	65	49	4default_deleteI
80	53	34	5F	45	45	4E	53	5F	39	61	6C	6C	6F	63	61	74	S4_EENS_9allocat
90	6F	72	49	53	34	5F	45	45	45	31	36	5F	5F	6F	6E	5F	orIS4_EEE16__on_
A0	7A	65	72	6F	5F	73	68	61	72	65	64	45	76	00	5F	5F	zero_sharedEv __
B0	5A	4E	53	74	33	5F	5F	31	32	30	5F	5F	73	68	61	72	ZNSt3__120__shar
C0	65	64	5F	70	74	72	5F	70	6F	69	6E	74	65	72	49	50	ed_ptr_pointerIP
D0	4E	33	70	63	69	39	69	6E	74	65	72	66	61	63	65	37	N3pci9interface7
E0	63	6F	6E	74	72	6F	6C	31	32	63	6C	69	65	6E	74	42	control2clientB
F0	75	66	66	65	72	45	4E	53	5F	31	34	64	65	66	61	75	ufferENS_14defau

Fig. 8 Data acquired from iPhone 5c NAND via reading software [76]

Zaharis et al. [95] propose an architecture which provides remote live forensics protection and eliminates malicious code execution in WSNs using sandboxing methods. Using the proposed architecture, one may dump the volatile memory from the sensor device. However, this architecture does not provide full memory dump for analysis, instead it extracts data selectively due to power efficiency constraints. The collected data is only used for verification of the integrity of the program that each sensor device is running. Nevertheless, this is not considered complete forensics analysis of sensor device memory.

In order to close the gap discussed above, Kumar et al. [45] propose an architecture of memory extraction from devices that are used in both IoT and WSNs environments. The main goal of this work is to investigate the extracted data in order to determine the reasons which could have caused the security breaches. This

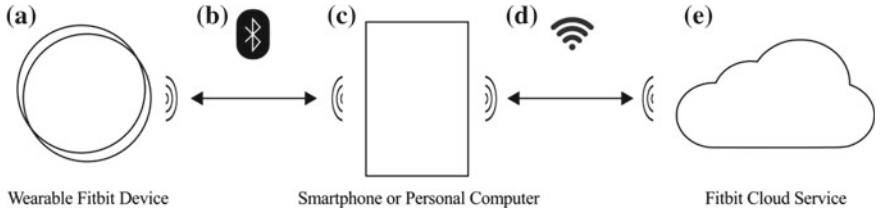


Fig. 9 The Fitbit system components: attack surface is partitioned into five regions, **a–e** [21]

architecture is specifically designed to extract, analyze, and correlate forensic data for IPv6-based WNS devices, which run Contiki [25] operating system and is powered by 8051-based, 8-bit microcontrollers. Contiki is a lightweight and open-source operating system for IoT and WSN devices. It is important to note that the analysis done by the authors is purely hardware based and does not depend on WSN traffic analysis.

This work is divided into three steps which are extraction, analysis, and co-relation. In the first step, a copy of memory is extracted from the device memory. In the second step, the acquired data is analyzed in a fully automated fashion in order to reduce the investigation time. In the final step, a set of new data is looked for by co-relating retrieved data from one device to another device in the case of multiple devices being used in the network.

As wearable IoT devices are becoming part of our everyday life, especially fitness trackers, they started to appear in the crime/incident scene and also being used in court cases [4, 59, 74]. This resulted in the need for forensic data collection from fitness trackers with different interfaces. Cyr et al. [21] have studied security analysis of Fitbit, a wearable fitness device. Although they mostly focused on the security issues in both device communication and mobile application, its importance is also negligible from the digital forensics perspective. This is mainly because their methods can be used by forensic investigators.

Figure 9 shows each component in Fitbit system when synchronization is performed between a Fitbit device, mobile device or computer, and Fitbit cloud service. This system is also partitioned into possible attack surfaces in the figure. In addition to security analysis, the same partitioning can also be used for forensic investigation as well. The device's memory can be extracted from Fig. 9a and analyzed using JTAG or chip-off techniques, and a chip reading software. Figure 9b, c can be attacked to read communication between both devices shown in Fig. 9a, c. Fitbit cloud data, however, can be retrieved using similar methods discussed later in Sect. 4.4.

In most of the wearable fitness devices, memory is packaged with waterproof material. Therefore, it is impossible to physically access the memory without destroying the packaging (see Fig. 10). Once the memory device is accessed, then JTAG or chip-off can be used to retrieve raw data from the memory.

Fig. 10 Fitbit Flex teardown process [34]

4.3 Network-Level Investigation

In some applications, IoT devices or sensors form a network of collective sensing and action. Therefore, in addition to device- level data, there will be data collected at the network level regarding the flow of data, routing, and tracking of lost packets. This IoT-related network may utilize one or more of the following networks:

- Body Area Network (BAN),
- Personal Area Network (PAN),
- Home/Hospital Area Networks (HAN),
- Local Area Networks (LAN),
- Wide Area Networks (WAN),
- Cyber-Physical System (CPS).

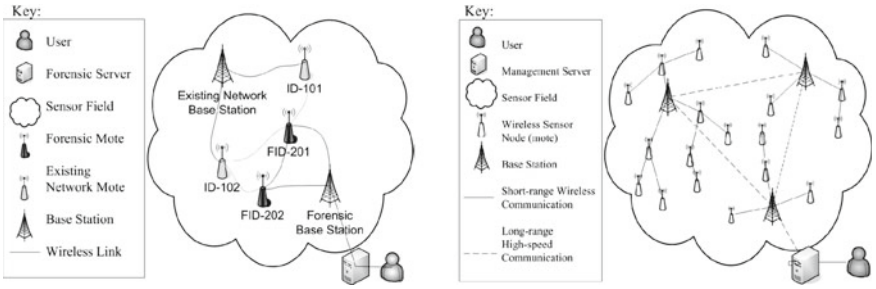
For each type of network, there needs to be customized mechanisms to be able to conduct cyber forensics after an incident. This forms a new form of research area that is different from the existing traditional wired networks.

Regardless of which form of network is used, most of the data in networks are volatile, and volatility of data causes serious issues in forensic investigations. Most of the hardware used in networks record transmitted data itself or some other information about that data in logs. These logs are indispensable to the forensic investigators as they may contain information which can eventually be used as evidence.

Firewalls capture and record the information about network traffic and keep the logs of events and transmitted data, which goes through them while preventing unauthorized access to the systems. Jahanbin et al. [36] proposed a design of autonomous intelligent multi-agent system in order to collect, examine, and analyze firewall logs, and report possible evidence related to an ongoing or previous criminal activities in WSNs.

The proposed architecture is designed to be located between the firewall and the end user, and it consists of three cognitive agents.

- *The collector agent*: This agent is used in collection step and responsible for collection and processing of the firewall logs that are recorded for a given WSN.



(a) A graphical representation of a wireless sensor network [55]

(b) A graphical representation of the network layout with digital forensics readiness implemented [56]

Fig. 11 Adding a digital forensics readiness layer to an existing Wireless Sensor Network for digital evidence collection

- *The inspector agent*: This agent is used in the inspection step and is responsible for identification of suspicious events from the given log files. It is also responsible for transmission of suspicious events to the next agent.
- *The investigator agent*: This agent is used in both investigation and notification steps. In the investigation step, it examines the forwarded suspicious event by the inspector agent and evaluates its effects and importance. It eventually decides whether it is malicious or not. In the notification step, the decisions are reported as security alerts to the security administrator in details.

It must be noted that, in order to preserve forensic soundness, the firewall logs must be checked for integrity purposes as users (either an administrator or adversary) might alter the logs and destroy the evidence (intentionally or unintentionally). All the agents mentioned the above work on the exact copy of the firewall log files and keep the originals as evidence in order to preserve integrity and provide reproducibility of forensic evidence.

Although WSNs have received the attention of security researchers, digital forensics research is still lacking in the discipline. In order to at least prepare WSNs for forensics investigations, Mouton and Venter [56] proposed a digital forensics readiness prototype in IEEE 802.15.4 WSNs. This prototype is designed based on the description made by Tan [84], who defines two digital forensics readiness objectives as follows:

1. Maximizing an environment's ability to collect credible digital evidence, and;
2. Minimizing the cost of forensics in an incident response.

Although Tan [84]'s objectives are sufficient enough for general digital forensics investigations, Mouton and Venter modified these objectives to be better suited to WSNs. Their objectives are threefold and aim to perform the investigation in the shortest amount of time, spending the least amount of time, and without causing disruptions in the network which may perform mission-critical tasks Mouton and Venter

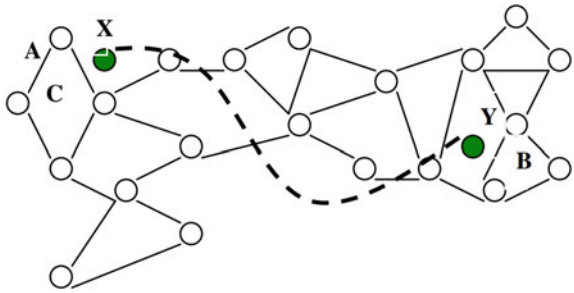


Fig. 12 Wormhole attack [7]

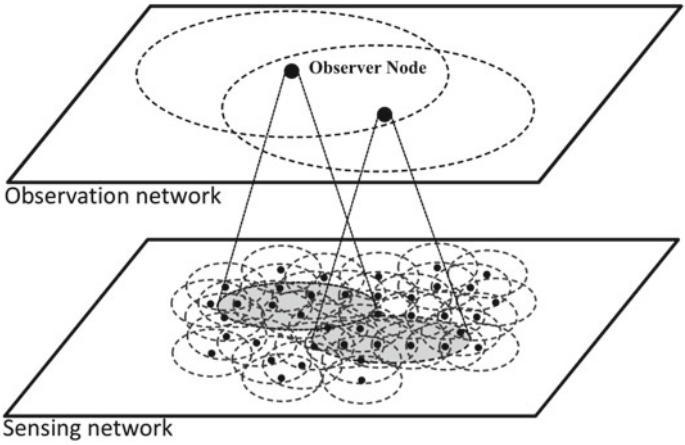


Fig. 13 Proposed WSN architecture [87]

[56]. As digital forensics investigations require the original source of evidence being protected against alterations, the last objective is critical for forensic soundness. In order to avoid inadmissibility of evidence and make the implementation of digital forensics readiness, the authors created additional independent forensics WSN referred as *fWSN* (see Fig. 11(a)), along with the original WSN referred as *oWSN* (see Fig. 11(b)).

Mouton and Venter also discuss the list of requirements (see Table 5), which can be used as a preliminary approach and need to be considered in order to implement digital forensics readiness in an IEEE 802.15.4 WSN environment. Note that the first column in the table shows some important factors, which make WNSs environments unique and different from WLAN.

In another work, [87] propose a solution to digital forensics investigations when wormhole attacks take place in a WSN. This solution ultimately aims to collect digital evidence, detect colluded nodes, and reconstruct the events which occurred during the wormhole attack which allows attackers to transmit a packet from one point to another point in the network by creating "tunnels" (see Fig. 12). This will

Table 5 Requirements in order to achieve digital forensic readiness in a IEEE 802.15.4 WSN environment [57]

Factors	Detailed requirement list
Communication protocol	1. The *fWSN* ensures the collection of all data packets by motes in the field utilizing a receipt acknowledgement packet protocol
	2. In order to make sure that the data packets are not changed, *oWSN*'s broadcasting communication should be intercepted
	3. All p*oWSN* ossible communication that is originating from *oWSN*
Proof of Authenticity and Integrity	4. While *fWSN* captures the data, the authenticity and integrity of all the data packets should be preserved
	5. Authenticity and integrity of the captured data in the *fWSN* should be preserved while they are being stored
	6. Verification on the authenticity and integrity of all the data packets should be available when digital investigation takes place
Time stamping	7. The data packets should have a time stamp assigned to them in order to preserve their authenticity and integrity
	8. The order of the captured packets should reflect the correct sequence when compared to the data transmitted from the original network
Modification of the network after deployment	9. It should be possible to implement the *fWSN* without any alteration in the *oWSN*
Protocol data packets	10. *fWSN*s operation should not be affected by the routing protocol or the network topology being used by *oWSN*
Radio frequencies	11. The *fWSN* should be able to communicate on the same radio frequencies that are available to the *oWSN*
	12. All communication within the *fWSN* should occur on a frequency not utilized in the *oWSN*
	13. Data packet should be captured forensically by the *fWSN* when an intruder WSN is in the area and communicates on a frequency that influences the *oWSN*

(continued)

Table 5 (continued)

Factors	Detailed requirement list
Power supply	14. In order to ensure that the *fWSN* captures all forensically relevant packets, the *fWSN* should have at least the same or a longer network lifetime than the *oWSN* in terms of battery power. Also, the *fWSN* should not increase power consumption in the *oWSN*
Network overhead	15. While intercepting communication, the *oWSN* should be free of extra network overhead
Data integrity	16. The *fWSN* should by no means be able to influence the *oWSN* or influence any sensory data transmitted within the *oWSN*

allow attackers to distribute the packet to other nodes from the second point in the network (see [7] for attack details).

The proposed solution suggests creating a virtual network called observation network, which consists of a set of investigator nodes and base stations. The nodes in this secondary network are called *observers*. Each observer has a limited coverage and they are responsible for monitoring the communication between observed nodes located in the sensing network (see architecture in Fig. 13). *Observers* collect information about the suspicious nodes such as traffic between nodes, routing path of data packets, and identity of those nodes. The aggregated evidence data is then broadcast to base stations.

Base stations are responsible for several activities which are defined in the following two groups [87].

1. Sensing-based activities such as:
   - Analyzing collected data from sensors
   - Creating decisions by correlating and filtering the data
   - Transmitting configuration to sensors
   - Activating sensors dynamically to reduce battery usage.
2. Investigation-based activities such as:
   - Collecting forensic information about the observed nodes from observation network
   - Analyzing, correlating, and merging the evidence in order to determine malicious nodes and rebuild the attack scenarios
   - Communicating with observers about configuration of their locations.

This proposed architecture is able to detect all types of wormhole attacks as the observer nodes and the observation network is designed in such a way that no observed nodes are left behind in the sensing network. In other words, all the

observed nodes are clustered into groups and each cluster is constantly observed by observer nodes.

Cyber Forensics for CPS and SCADA. Recently, IoT has also started to be deployed in control systems for actuation purposes, which led to the concept of CPS [40, 68]. In such systems, IoT devices are involved in sensing, communicating, and acting. The difference from the above networks is that there are nodes which do actuation and thus this creates another venue for forensics investigation. One form of CPS is in the area of control systems for critical applications such as energy, transportation and industry. In such systems, the network is referred to as Supervisory Control and Data Acquisition (SCADA) [14] and failure or attacks in such systems is crucial to be detected and investigated for the applications to sustain [44]. SCADA systems, are used for the collection and analysis of real-time data from Industrial Control Systems (ICS). Most of the CPSes rely on computer and control systems in order to provide reliable operations to safeguard the infrastructure. Therefore, forensic analysis of SCADA/ICS systems has been an important tool which was considered in some works. In the remainder of this section, we also discuss these approaches as they relate to a network-level investigation.

SCADA systems consist of a field site and control center. In the field site, there are IoT devices which are considered as intelligent such as Programmable Logic Controllers (PLCs) or Remote Terminal Units (RSUs). These are typically attached to physical processes such as thermostats, motors, and switches. The control center is responsible for collecting data related to the state of field instruments and interacting with the field sites. Components found at the control center typically consist of a Human Machine Interface (HMI), Historian and Master Terminal Unit (MTU). All of these are connected with a LAN that can run various protocols including MODBUS [54], DNP3 [18], and Ethernet.

The information security vulnerabilities of ICS have been studied extensively, and the vulnerable nature of these systems is well known [63, 71, 79]. However, in the case of a security incident (e.g., denial of service attacks), it is important to understand what the digital forensics consequences of such an attack are? What procedures or protocols are needed to be used during an investigation? What tools and techniques are appropriate to be used by the investigator? Where can forensic data be collected and how?

In the rest of this section, we discuss various research efforts aimed at assisting SCADA forensics procedures by proposing tools, techniques, and forensics investigation models.

SCADA Live Forensics: SCADA is originally deployed to non-networked environments, therefore there has been a lack of security against Internet-based threats and cyber-related forensics. Over time, there has been a huge increase in the vulnerability of threats caused through connectivity allowing remote control over the Internet. The attacks necessitated SCADA system a forensic investigation in order to understand the effects and cause of the intrusion. Taveras [85] focuses on detecting the abnormal changes of sensor reads, illegal penetrations, failures, traffic over the communication channel, and physical memory content by creating a software application. The

challenging issue is that the tool should be developed in a way that it should have the minimal impact on the SCADA resources during the data acquisition process.

The problem involved in this process is that SCADA systems should not be turned off for data acquisition and analysis as it is being continuously operational. There has not been a single forensic tool to preserve the hardware and software state of a system during investigation. Research continued to provide a computing module to support the incident response and digital evidence collection process. Experiment is performed on the SCADA system by performing live data acquisition and then performing subsequent offline analysis of the acquired data.

Based on the live forensic analysis of the data collected from the SCADA system, it is concluded that traditional information security mechanisms cannot be applied directly as these systems cannot tolerate delays in performance which eventually require a lot of memory to perform long processes. Thus, SCADA systems should consider a special operating paradigm. This also paved the way to improve the infrastructure of the systems and provide appropriate tools for forensic analysis over interconnected SCADA systems.

Limitations of forensic analysis tools on SCADA systems: As Ahmed et al. [1] notes, currently available traditional digital forensics tools are not capable of performing data analysis on SCADA systems. The main reason is that state-of-the-art tools are designed to work on deterministic systems and devices such as hard disk drives, mobile phones, network traffic captures saved in pcap files. However, SCADA systems generate propriety log data depending on the make and model the hardware. As discussed above, investigators are in need of creating new scripts for their own particular needs to overcome this issue. Hence, there is an expectation from the research community and forensics tools developers to design SCADA forensics tools or patch currently available tools in order to respond to this demand.

Developing Forensics Investigation Models for SCADA: Once the vulnerabilities and the possible attacks on the SCADA/ICS systems are analyzed, it is crucial to perform forensically sound forensic analysis on SCADA/ICS. The current literature shows some efforts of developing forensic analysis frameworks and models.

One of the early frameworks is proposed by Wu et al. [92]. In addition to this framework, Stirland et al. [78] proposed a methodology to analyze the problems involved in SCADA/ICS systems proposed (see Fig. 14). The authors in this work particularly categorize a set of forensic toolkits (both commercial and open source) to support each stage of an investigation and structure of the control systems. The proposed methodology involves a clear process of investigation, which includes the following phases:

1. Identification and preparation of the requirements and the problem involved in extracting the evidence.
2. Identifying data sources—this phase involves gathering the data from sources and analyzing if the system supports the data sources.
3. Preservation, prioritizing, and collection this phase works depending on the priority of data and different data capturing techniques are involved to ensure all devices are captured or not.

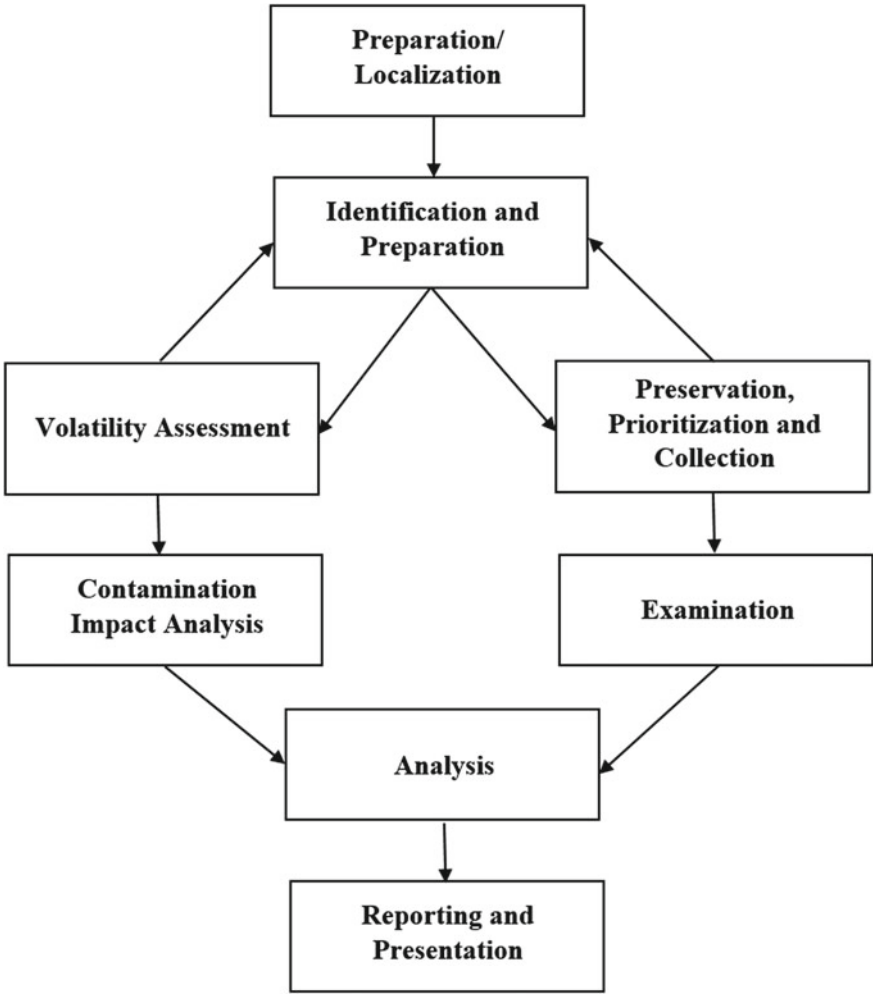


Fig. 14 Cyber forensics model for SCADA/ICS by Stirland et al. [78]

4. Examination and analysis—this phase involves in performing the analysis depending on the data sources, methods and provides a timeline in preparing the data and logs on it and allows to extract data.
5. Reporting and presentation- this phase involves in providing a report to all the details performed in the above phases including the outcome of the analysis which also includes documentation of the further recommendations for future study.

Security is of high importance for the control systems and there are many recommendations for further improvements in incident response to support investigation

Table 6 Forensic toolkit application to SCADA systems

SCADA device	Phase	Forensic tool
Network	Phase 3	TCPDump
	Phase 4	Network Miner, Wireshark, AlienVault
HMI	Phase 3	Write Blockers, FTK Imager, EnCase, Helix, SHA-256/MD5 Hashing tool
	Phase 4	EnCase, XWays, Accessdata FTK, Volatility
PLC/RTU	Phase 3	Besope PLC Flashing Software
	Phase 4	XWays
Engineering computer	Phase 3	Write blockers, FTK Imager, EnCase, Helix, SHA-256/MD5 Hashing tool
	Phase 4	EnCase, XWays, Accessdata FTK, Volatility
Database server	Phase 3	Write Blockers, FTK Imager, EnCase, Helix, SHA-256/MD5 Hashing tool
	Phase 4	EnCase, XWays, Accessdata FTK, Volatility
OPC	Phase 3	Write Blockers, FTK Imager, EnCase, Helix, SHA-256/MD5 Hashing tool
	Phase 4	EnCase, XWays, Accessdata FTK, Volatility
Historian	Phase 3	Write Blockers, FTK Imager, EnCase, Helix, SHA-256/MD5 Hashing tool
	Phase 4	EnCase, XWays, Accessdata FTK, Volatility

and increase the level of complexity for attacking the systems by attackers. It is concluded that the proposed methodology for developing a forensics toolkit is considered based on the requirements of SCADA systems. Various suggested tools are shown in Table 6. There are already existing tools, which support the elements of SCADA forensic investigation and further research and progress in this area is needed in order to identify more evidence and artifacts.

Moreover, Ahmed et al. [1] discuss the challenges for forensics investigators in SCADA systems and their potential solutions. In order to address some of these unique challenges of SCADA forensics analysis, a recent framework was proposed in Eden et al. [26]. This framework aims at identifying necessary steps for incident response as well. Figure 15 shows the SCADA forensic incident response model consisting of six main stages: (1) Prepare; (2) Detect; (3) Isolate; (4) Triage; (5) Respond; and (6) Report.

This proposed model suggests that the preparation phase must be performed prior to incident the happening. In this first stage, an investigator must understand the system's architecture with respect to its configuration and hardware devices used in the system. This step is crucial to the first responders to avoid complication, when the recovery from an incident is time sensitive. In this stage, it is made sure that all the hardware used in the system is well documented. In addition to the system architecture, the forensics investigator is also expected to know the SCADA system's requirements with respect to availability of the system. This is essential

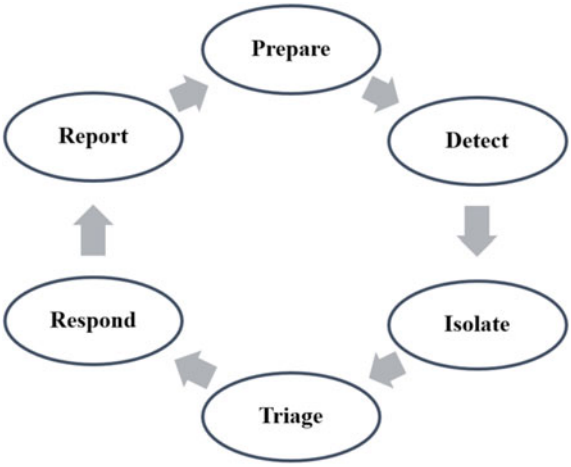


Fig. 15 Forensics model for SCADA systems

particularly if the system is mission critical and running states of a certain device must be preserved while the investigation is in progress. Therefore, prior knowledge of system requirements plays a critical role. Finally, the investigator is also expected to understand potential attacks targeting hardware, software, and the communication stack of the SCADA system. Such knowledge will allow the investigators to better perform in the following phases.

In the second stage of the model, the investigator is expected to detect the type of attack and potential infected areas in the system. This detection will be performed based on the real-time data available in the system such as network packages and log files. Once the type of attack has been determined, investigators will be able to locate infected areas based on the behavior of the attack. As long as the infected areas are detected, it will be easier to know the type of data in the next stage.

The isolate stage is critical for the investigation with respect to the importance of SCADA system in the CPS. In most of cases, infected areas in the network must be isolated so that further contamination and disruption to the system can be avoided. The success of the isolation will be dependent on the success of the detection of potential infected areas.

Despite the classical forensics investigations, triaging is different in SCADA networks. Forensics investigators must identify the data sources in order to triage the available data. This will also be dependent on the information (e.g., device make, model, and serial number) provided in earlier stages, particularly preparation and detection stages. Once the data identification is performed, then data sources ought to be prioritized with respect to the value, volatility, and accessibility of data. This will allow investigators to acquire as much evidence as possible.

In the respond stage, investigators perform actual data acquisition from the SCADA system (network) by using the priority list created in the previous stage. As a rule of thumb, data must be acquired from the SCADA system by using forensically sound tools and techniques. As discussed before, this will prove the admissibility

of evidence in a court of law (when needed). In order to acquire data from various devices and network, aforementioned forensics acquisition techniques can be used. Once the data acquisition is completed, then analysis of data is performed using available forensics tools or by creating new special scripts for unconventional data. Eventually, the aim is to find a forensics artifact to be presented as evidence from the large set of unrelated system data.

Finally, similar to the traditional investigations, the reporting stage requires investigators to document all the steps taken, tools used, evidence collected, and challenges faced. When they are documented systematically, then the investigator may create a timeline of events by reviewing the findings to support the evidence found and determine the source of an incident/attack. The final report must comply with the chain of custody by providing validation and verification of evidence found.

Our final discussion in this section is briefly on accurate modeling of the SIEMENS S7 SCADA Protocol for intrusion detection and digital forensics using real-life data. Siemens S7 is used in SCADA systems for communications between a HMI and the Programmable Logic Controllers (PLCs). In Kleinmann and Wool [41], Intrusion Detection system (IDS) model is designed for S7 networks which analyzes the traffic to and from a specific PLC. A unique Deterministic Finite Automata (DFA) is used to model the HMI-PLC channel traffic whether it is highly periodic or not.

SCADA systems have its own strategy in analyzing the fault or malfunction. In this paper, it is defined that the research based on traffic simulation has several risks such as lack of realism which effects the use of SCADA systems. Three different traces of datasets are collected in order to perform the experiments which are collected at ICS facilities. The first S7 SCADA trace was collected from a manufacturing plant, where a single channel is observed between the HMI and an S7 complaint VIPA PLC. Next, two traces are collected from a water treatment facility which has control over specific levels in tanks. A Wireshark program is used to collect the traces with HMI running in background in the operating system. The authors show that, based on the analysis of the traffic from two ICS plants, some key semantics of the proprietary of S7 protocol can be reverse engineered. It is also observed that previously developed Modbus showed successful results in the same way DFA-based approach is very successful with high accuracy and extremely low-false positive rates; IDs is further extremely efficient which works at line speed to detect the anomalies.

4.4 Cloud-Level Investigation

As discussed in the previous sections, forensic investigation in cloud environments has its own challenges such as multi-tenancy and multi-jurisdiction. Since IoT devices have limited storage and computational resources, the actual data is processed and stored in the cloud. This causes investigations being conducted in the cloud environment especially when data in physical storage and network does not result in useful evidence. Hence, similar investigative challenges in the cloud exist when forensic investigations in IoT are conducted. Although current research efforts in IoT foren-

sics are in their very early stages, there are some successful models suggesting easier investigations in the cloud environment. In this section, we will specifically focus on IoT forensics investigation models proposed for cloud environments.

According to Zawoad and Hasan [96], the term IoT forensics was not formally defined until they proposed forensics-aware model (see Fig. 16) for IoT infrastructures called FAIoT. This model supports digital evidence collection and analysis in the IoT environment by providing easiness and forensic soundness. Such a model might also allow cloud service providers addressing the needs of law enforcement officers when a search warrant is obtained to collect data from cloud environments.

Secure Evidence Preservation Module: This module provides constant monitoring of the registered devices for forensics evidence in the form of logs files or data collected by sensors. If evidence is recognized, it is then stored in the evidence repository. Evidence repository database is designed on top of Hadoop Distributed File System (HDFS) in order to provide scalable and reliable data processing of large data. The data kept in the database will be categorized based on the IoT device and its owner in order to reduce multi-tenancy issues and avoid commingling of data in cloud [96].

Secure Provenance Module: This module provides a chain of custody for the evidence stored and kept in the database. This is made possible by using Provenance-Aware File System (PASS) introduced by Muniswamy-Reddy et al. [58]. PASS is a storage system which performs automated collection, management, storage, and search of provenance an object [58]. Secure Provenance Module provides provenance record of evidence stored in provenance database by using PASS.

Finally, investigators can access the evidence and its provenance record using the proposed APIs which makes sure the confidentiality of evidence by using encryption algorithms. In order for this to be possible, investigators need a Web Server to access the requested data through the API.

In another work, Oriwoh and Sant [62] propose a more specialized model called Forensics Edge Management System (FEMS), which is specific to smart home environments. FEMS is an automated system which can be integrated into smart homes in order to perform initial forensic investigations while providing basic security services [62]. Figure 17 shows the architecture of the FEMS and all the security and forensic services provided by FEMS.

Oriwoh and Sant also proposed a digital forensics framework called IoT Digital Forensics Framework (IDFF) (see Fig. 18). This framework presents step-by-step operation presented in the flowchart in order to show how FEMS operation is performed. As stated by the authors, usage of FEMS provides automatic, intelligent, and autonomous detection and investigation, and indicates the source of security issues in smart homes to its users.

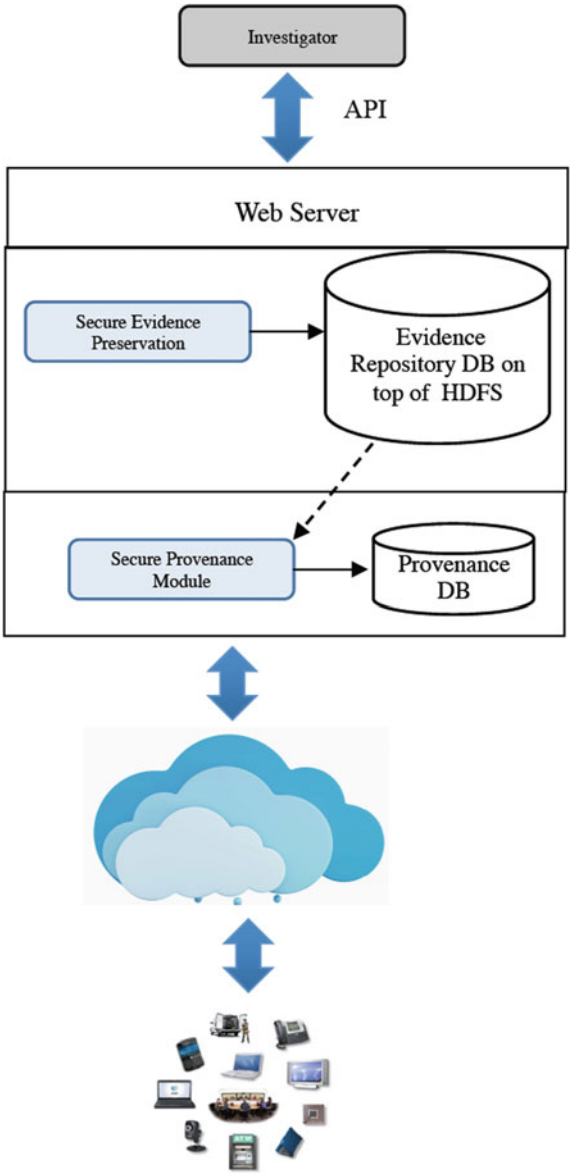


Fig. 16 The proposed model of IoT forensics [96]

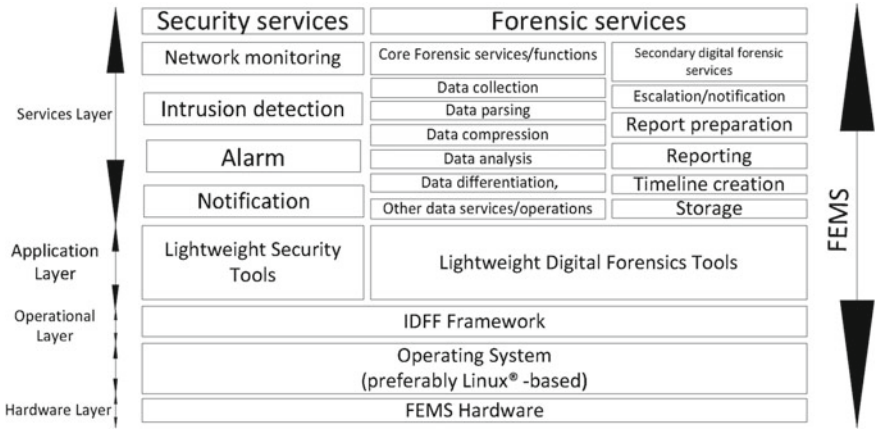


Fig. 17 The FEMS architecture [62]

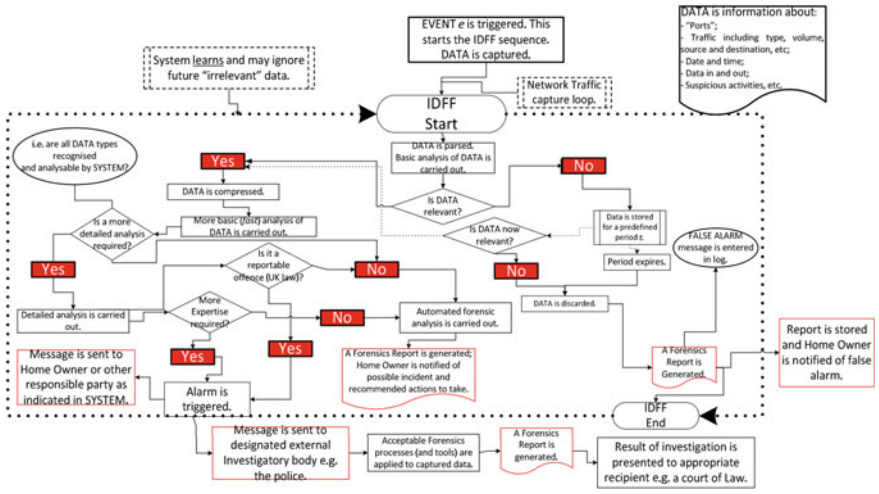


Fig. 18 Detailed flowchart of the IoT digital forensics framework (IDFF) [62]

4.5 Future Research

As digital forensics in IoT and WSNs is a relatively new concept particularly for the digital forensics community, the current research reveals important future work to be conducted. It is quite obvious that newer investigative techniques will be soon needed as Google has announced Android Things [29] a new operating system to develop new IoT devices and there is a growing number of IoT devices being deployed in our daily applications. It is worth noting that as WSNs are now widely considered under the IoT concept, most of the endeavors for digital forensics research will be

also applicable for WSNs. In addition, digital forensic solutions and frameworks also will be needed due to the availability of low memory footprint and low-power requirement devices used in WSNs.

Despite this inevitable demand in the very near future, once they are developed, these new methods will ultimately help investigators to perform standard investigative processes. Until then, it could also be possible to use some of the available tools and techniques that are readily available for current Android operating systems. In addition to the existing research literature, we believe some of the areas for further work could be listed as follows:

- Standardize data storage units and interfaces in similar devices. Forensically valuable IoT devices (e.g., fitness trackers) could be designed and manufactured with data storage units which can be analyzed using state-of-the-art forensic tools. Using known interfaces such as JTAG connections for IoT devices is also critical for faster and reliable investigations.
- Develop automated decision-making systems on forensically sound data for specific IoT technologies such as smart homes. It is well known that artificial intelligence techniques have been applied to many digital forensic domains to intelligently automate duties performed by human entities. Therefore, as an example, it can be very useful to adapt machine learning techniques to classify evidence in IoT domain or expert systems can be used to create intelligent tools to make decisions based on knowledge collected from both investigators and IoT environments.
- Build a model that would correlate evidence found in IoT environments. Digital forensics evidence correlation is an important concept especially when heterogeneous data is involved in investigations. Case et al. [15] have developed a framework for automatic evidence discovery and correlation from a variety of forensic targets. We also believe that similar models can also be built for IoT environments in order to use unrelated data leading to actual evidence through correlation.
- Analyze Android Things and develop new forensics models and tools for data acquisition, examination, analysis, and reporting. This brand new operating system needs immediate attention from the researchers as it is projected to be used in many IoT devices in the near future.
- Create new digital forensics investigation models (e.g., Electronic Discovery Reference Model, see http://www.edrm.net) for specific IoT environments. Due to the heterogeneity of data and hardware in IoT devices, it could be useful to develop IoT specific investigation models. Because, currently available models are mostly designed for storage, network, and cloud specific, however, IoT environments may necessitate all three environments being used.
- Collaborate with data analytics and fault-tolerance experts to cooperatively analyze data from IoT devices not only related to user activity but also related to hardware and embedded systems. This opens up opportunities for insurance companies as they would like to investigate issues regarding failures while some of these failures might be due to actual attacks from external attackers.

- Create robust and standard solutions particularly for live data acquisition, automated data collection, recovery of memory and processes from live units in SCADA systems.
- Develop legal solutions to the issues including preservation of the chain of custody and admissibility of IoT evidence. In digital investigations, it is critical to preserve chain of custody for evidence admissibility. However, it may not be possible in IoT environments because of their designs. Involvement in legislative processes regarding IoT forensics investigations is needed to determine solutions from the legal aspects.

4.6 Conclusion

The IoT and WSNs offer a significant source of potential evidence, however due to their heterogeneous nature and the ways in which data is distributed, aggregated, and processed, there are challenges that the digital forensics investigations must overcome. For this purpose, new techniques are required to not only overcome the hurdles, but also influence the architecture and processes in order to gain access to this rich source of potential evidence in the IoT and thus WSN environments. In this book chapter, we explained digital forensics challenges in IoT and WSN environments. We also analyzed and explained currently available solutions to overcome some of those challenges from different perspectives. As discussed in the Sect. 4.5, there are still many open research problems in this new area.

References

1. Ahmed, I., Obermeier, S., Naedele, M., Richard III, G.G.: Scada systems: challenges for forensic investigators. Computer **45**(12), 44–51 (2012)
2. Akkaya, K., Younis, M.: A survey on routing protocols for wireless sensor networks. Ad Hoc Netw. **3**(3), 325–349 (2005)
3. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: A survey on sensor networks. IEEE Commun. Mag. **40**(8), 102–114 (2002)
4. Alba, A.: Police, attorneys are using fitness trackers as court evidence (2016). http://www.nydailynews.com/news/national/police-attorneys-fitness-trackers-court-evidence-article-1.2607432
5. Anderson, D.: White paper: red hat crash utility (2008)
6. Arora, A., Dutta, P., Bapat, S., Kulathumani, V., Zhang, H., Naik, V., Mittal, V., Cao, H., Demirbas, M., Gouda, M., et al.: A line in the sand: a wireless sensor network for target detection, classification, and tracking. Comput. Netw. **46**(5), 605–634 (2004)
7. Arora, M., Challa, R.K., Bansal, D.: Performance evaluation of routing protocols based on wormhole attack in wireless mesh networks. In: Second International Conference on Computer and Network Technology, pp. 102–104. IEEE (2010)
8. Ashton, K.: That internet of things thing. RFiD J. **22**(7), 97–114 (2009)
9. Atzori, L., Iera, A., Morabito, G.: The internet of things: a survey. Comput. Netw. **54**(15), 2787–2805 (2010)

10. Ayers, R., Brothers, S., Jansen, W.: Guidelines on Mobile Device Forensics (draft), vol. 800, p. 101. NIST Special Publication (2013)
11. Baccelli, E., Hahm, O., Gunes, M., Wahlisch, M., Schmidt, T.C.: Riot os: towards an os for the internet of things. In: 2013 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 79–80. IEEE (2013)
12. Baronti, P., Pillai, P., Chook, V.W., Chessa, S., Gotta, A., Hu, Y.F.: Wireless sensor networks: A survey on the state of the art and the 802.15. 4 and zigbee standards. Comput. Commun. **30**(7), 1655–1695 (2007)
13. S. Bluetooth: Bluetooth specification version 1.1 (2001). http://www.bluetooth.com
14. Boyer, S.A.: SCADA: Supervisory Control and Data Acquisition. International Society of Automation (2009)
15. Case, A., Cristina, A., Marziale, L., Richard, G.G., Roussev, V.: Face: automated digital evidence discovery and correlation. Digital Invest. **5**, S65–S75 (2008)
16. Chen, Y.-K.: Challenges and opportunities of internet of things. In: 17th Asia and South Pacific Design Automation Conference, pp. 383–388. IEEE (2012)
17. Christin, D., Reinhardt, A., Mogre, P.S., Steinmetz, R., et al.: Wireless sensor networks and the internet of things: selected challenges. In: Proceedings of the 8th GI/ITG KuVS Fachgespräch Drahtlose sensornetze, pp. 31–34 (2009)
18. Clarke, G.R., Reynders, D., Wright, E.: Practical modern SCADA protocols: DNP3, 60870.5 and related systems, Newnes (2004)
19. Committee, I.S., et al.: Isa100. 11a, wireless systems for industrial automation: process control and related applications. Technical Report, Research Triangle Park, North Carolina (2009)
20. C. Computing: Toward a multi-tenancy authorization system for cloud services (2010)
21. Cyr, B., Horn, W., Miao, D., Specter, M.: Security analysis of wearable fitness devices (fitbit), p. 1. Massachusets Institute of Technology (2014)
22. Demirkol, I., Ersoy, C., Alagoz, F.: Mac protocols for wireless sensor networks: a survey. IEEE Commun. Mag. **44**(4), 115–121 (2006)
23. I. Devices. Various iot devices (2016). http://iotlist.co
24. DoJ: Electronic crime scene investigation: a guide for first responders (2001)
25. Dunkels, A., Gronvall, B., Voigt, T.: Contiki-a lightweight and flexible operating system for tiny networked sensors. In: 2004 29th Annual IEEE International Conference on Local Computer Networks, pp. 455–462. IEEE (2004)
26. Eden, P., Blyth, A., Burnap, P., Cherdantseva, Y., Jones, K., Soulsby, H., Stoddart, K.: A Cyber Forensic Taxonomy for SCADA Systems in Critical Infrastructure, pp. 27–39. Springer International Publishing, Cham (2016). https://doi.org/10.1007/978-3-319-33331-1_3. ISBN 978-3-319-33331-1
27. Estrin, D., Govindan, R., Heidemann, J., Kumar, S.: Next century challenges: scalable coordination in sensor networks. In: Proceedings of the 5th annual ACM/IEEE International Conference on Mobile Computing and Networking, pp. 263–270. ACM (1999)
28. Gaura, E., Newman, R.: Smart MEMS and Sensor Systems. World Scientific (2006)
29. Google: Android things (2016). https://developer.android.com/things/index.html
30. Google: Google trends (2016). https://www.google.com/trends
31. Gubbi, J., Buyya, R., Marusic, S., Palaniswami, M.: Internet of things (iot): a vision, architectural elements, and future directions. Future Gener. Comput. Syst. **29**(7), 1645–1660 (2013)
32. Hegarty, R., Lamb, D., Attwood, A.: Digital evidence challenges in the internet of things. In: Proceedings of the Tenth International Network Conference (INC 2014), p. 163 (2014). www.Lulu.com
33. Hosain, S.: Reality check: 50b iot devices connected by 2020 beyond the hype and into reality (2016). http://www.rcrwireless.com/20160628/opinion/reality-check-50b-iot-devices-connected-2020-beyond-hype-reality
34. iFixit. Fitbit flex teardown (2016). https://www.ifixit.com/Teardown/Fitbit+Flex+Teardown/16050
35. B. Intelligence: Chip-off forensics (2016). http://www.binaryintel.com/services/jtag-chip-off-forensics/chip-off_forensics/

36. Jahanbin, A., Ghafarian, A., Seno, S.A.H., Nikookar, S.: Computer forensics approach based on autonomous intelligent multi-agent system. Int. J. Database Theory Appl. **6**(5), 1–12 (2013)
37. Jiang, L., Da Xu, L., Cai, H., Jiang, Z., Bu, F., Xu, B.: An iot-oriented data storage framework in cloud computing platform. IEEE Trans. Ind. Inform. **10**(2), 1443–1451 (2014)
38. Jiang, Y., Zhang, L., Wang, L.: Wireless sensor networks and the internet of things. Int. J. Distrib. Sens. Netw. **9**(6), 589750 (2013). https://doi.org/10.1155/2013/589750
39. Karabiyik, U.: Building an intelligent assistant for digital forensics. Ph.D Thesis, The Florida State University (2015)
40. Khaitan, S.K., McCalley, J.D.: Design techniques and applications of cyberphysical systems: a survey. IEEE Syst. J. **9**(2), 350–365 (2015)
41. Kleinmann, A., Wool, A.: Accurate modeling of the siemens s7 scada protocol for intrusion detection and digital forensics. J. Digit. Forensics Secur. Law **9**(2), 4 (2014)
42. Ko, J., Terzis, A., Dawson-Haggerty, S., Culler, D.E., Hui, J.W., Levis, P.: Connecting low-power and lossy networks to the internet. IEEE Commun. Mag. **49**(4) (2011)
43. Kollár, I.: Forensic Ram Dump Image Analyzer (2010)
44. Krutz, R.L.: Securing SCADA Systems. Wiley (2005)
45. Kumar, V., Oikonomou, G., Tryfonas, T., Page, D., Phillips, I.: Digital investigations for ipv6-based wireless sensor networks. Digit. Investig. **11**, S66–S75 (2014)
46. Kumar, V., Oikonomou, G., Tryfonas, T.: Traffic forensics for ipv6-based wireless sensor networks and the internet of things. In: 2016 IEEE 3rd World Forum on Internet of Things (WF-IoT), pp. 633–638. IEEE (2016)
47. Levis, P., Madden, S., Polastre, J., Szewczyk, R., Whitehouse, K., Woo, A., Gay, D., Hill, J., Welsh, M., Brewer, E., et al.: Tinyos: an operating system for sensor networks. Ambient Intell. **35**, 115–148 (2005)
48. Li, F., Xiong, P.: Practical secure communication for integrating wireless sensor networks into the internet of things. IEEE Sens. J. **13**(10), 3677–3684 (2013)
49. Li, T., Liu, Y., Tian, Y., Shen, S., Mao, W.: A storage solution for massive iot data based on nosql. In: 2012 IEEE International Conference on Green Computing and Communications (GreenCom), pp. 50–57. IEEE (2012)
50. Mainetti, L., Patrono, L., Vilei, A.: Evolution of wireless sensor networks towards the internet of things: a survey. In: 2011 19th International Conference on Software, Telecommunications and Computer Networks (SoftCOM), pp. 1–6. IEEE (2011)
51. Manjrasoft: Aneka: enabling .net-based enterprise grid and cloud computing (2016). http://www.manjrasoft.com
52. Minerva, R., Biru, A., Rotondi, D.: Towards a definition of the internet of things (iot). IEEE Internet Initiative 1 (2015)
53. Miorandi, D., Sicari, S., De Pellegrini, F., Chlamtac, I.: Internet of things: vision, applications and research challenges. Ad Hoc Netw. **10**(7), 1497–1516 (2012)
54. I. Modbus: Modbus application protocol specification v1. 1a. North Grafton, Massachusetts (2004). www.modbus.org/specs.php
55. Mouton, F., Venter, H.: A secure communication protocol for wireless sensor networks. In: Proceedings of the Annual Security Conference, Security Assurance and Privacy: Organizational Challenges, Las Vegas (2009)
56. Mouton, F., Venter, H.: A prototype for achieving digital forensic readiness on wireless sensor networks. In: AFRICON, pp. 1–6. IEEE (2011)
57. Mouton, F., Venter, H.S.: Requirements for wireless sensor networks in order to achieve digital forensic readiness. In: WDFIA, pp. 108–121 (2011)
58. Muniswamy-Reddy, K.-K., Holland, D.A., Braun, U., Seltzer, M.I.: Provenance-aware storage systems. In: USENIX Annual Technical Conference, General Track, pp. 43–56 (2006)
59. News4JAX: Fitness tracker data used in court cases (2016). http://www.click2houston.com/news/fitness-tracker-data-used-in-court-cases
60. Nikkel, B.J.: An introduction to investigating ipv6 networks. Digit. Invest. **4**(2), 59–67 (2007)
61. OpenEmbedded: Openembedded, the build framework for embedded linux (2017). http://www.openembedded.org/wiki/Main_Page

62. Oriwoh, E.,Sant, P.: The forensics edge management system: a concept and design. In: Ubiquitous Intelligence and Computing, 2013 IEEE 10th International Conference on and 10th International Conference on Autonomic and Trusted Computing (UIC/ATC), pp. 544–550. IEEE (2013)
63. Patzlaff, H.: D7. 1 preliminary report on forensic analysis for industrial systems. In: CRISALIS Consortium, Symantec, Sophia Antipolis, France (2013)
64. Peña-López, I.: ITU internet report 2005: the internet of things (2005)
65. Pereira, P.P., Eliasson, J., Kyusakov, R., Delsing, J., Raayatinezhad, A., Johansson, M.: Enabling cloud connectivity for mobile internet of things applications. In: 2013 IEEE 7th International Symposium on Service Oriented System Engineering (SOSE), pp. 518–526. IEEE (2013)
66. Perumal, S., Norwawi, N.M., Raman, V.: Internet of things (iot) digital forensic investigation model: top-down forensic approach methodology. In: 2015 Fifth International Conference on Digital Information Processing and Communications (ICDIPC), pp. 19–23. IEEE (2015)
67. Rabieh, K., Akkaya, K., Karabiyik, U., Qamruddin, J.: A secure and cloud-based medical records access scheme for on-road emergencies. In: 2018 15th IEEE Annual Consumer Communications & Networking Conference (CCNC), pp. 1–8. IEEE (2018)
68. Rajkumar, R.R., Lee, I., Sha, L., Stankovic, J.: Cyber-physical systems: the next computing revolution. In: Proceedings of the 47th Design Automation Conference, pp. 731–736. ACM (2010)
69. Ruan, K., Carthy, J., Kechadi, T., Crosbie, M.: Cloud forensics. In: IFIP International Conference on Digital Forensics, pp. 35–46. Springer (2011)
70. Saltaformaggio, B., Bhatia, R., Zhang, X., Xu, D., Richard III, G.G.: Screen after previous screens: Spatial-temporal recreation of android app displays from memory images. In: USENIX Security Symposium, pp. 1137–1151 (2016)
71. Shahzad, A., Musa, S., Aborujilah, A., Irfan, M.: Industrial control systems (icss) vulnerabilities analysis and scada security enhancement using testbed encryption. In: Proceedings of the 8th International Conference on Ubiquitous Information Management and Communication, pp. 7. ACM (2014)
72. Shelby, Z., Bormann, C.: 6LoWPAN: the wireless embedded Internet, vol. 43. Wiley (2011)
73. Sheng, Z., Yang, S., Yu, Y., Vasilakos, A., Mccann, J., Leung, K.: A survey on the ietf protocol suite for the internet of things: standards, challenges, and opportunities. IEEE Wirel. Commun. **20**(6), 91–98 (2013)
74. Siegal, J.: One Womans Fitbit just Decided a Criminal Case (2016). http://bgr.com/2016/04/20/fitbit-fitness-tracker-legal-case/
75. SiliconLabs: Battery size matters (2016). http://www.silabs.com/products/wireless/Pages/battery-life-in-connected-wireless-iot-devices.aspx
76. Skorobogatov, S.: The bumpy road towards iphone 5c nand mirroring (2016). arXiv:1609.04327
77. Song, J., Han, S., Mok, A., Chen, D., Lucas, M., Nixon, M., Pratt, W.: Wirelesshart: applying wireless technology in real-time industrial process control. In: Real-Time and Embedded Technology and Applications Symposium, 2008. RTAS'08. IEEE, pp. 377–386. IEEE (2008)
78. Stirland, J., Jones, K., Janicke, H., Wu, T.: Developing cyber forensics for scada industrial control systems. In: The International Conference on Information Security and Cyber Forensics (InfoSec2014), pp. 98–111. The Society of Digital Information and Wireless Communication (2014)
79. Stouffer, K., Falco, J., Scarfone, K.: Guide to Industrial Control Systems (ICS) Security, vol. 800, no. 82, pp. 16–16. NIST Special Publication (2011)
80. Sundmaeker, H., Guillemin, P., Friess, P., Woelfflé, S.: Vision and challenges for realising the internet of things. Cluster of European Research Projects on the Internet of Things, European Commision (2010)
81. Swauger, J.: Chip-off Forensics (2012)
82. Sylve, J., Case, A., Marziale, L., Richard, G.G.: Acquisition and analysis of volatile memory from android devices. Digit. Invest. **8**(3), 175–184 (2012)
83. Szewczyk, R., Mainwaring, A., Polastre, J., Anderson, J., Culler, D.: An analysis of a large scale habitat monitoring application. In: Proceedings of the 2nd International Conference on Embedded Networked Sensor Systems, pp. 214–226. ACM (2004)

84. Tan, J.: Forensic Readiness. Cambridge, MA, Stake, pp. 1–23 (2001)
85. Taveras, P.: Scada live forensics: real time data acquisition process to detect, prevent or evaluate critical situations. Eur. Sci. J. ESJ, **9**(21) (2013)
86. Tillman, K.: How many internet connections are in the world? right now (2013). http://blogs.cisco.com/news/cisco-connections-counter
87. Triki, B., Rekhis, S., Boudriga, N.: Digital investigation of wormhole attacks in wireless sensor networks. In: 2009 Eighth IEEE International Symposium on Network Computing and Applications, NCA 2009, pp. 179–186. IEEE (2009)
88. Walters, J.P., Liang, Z., Shi, W., Chaudhary, V.: Wireless sensor network security: a survey. In: Security in Distributed, Grid, Mobile, and Pervasive Computing vol. 1, p. 367 (2007)
89. Wang, C., Sohraby, K., Li, B., Daneshmand, M., Hu, Y.: A survey of transport protocols for wireless sensor networks. IEEE Netw. **20**(3), 34–40 (2006)
90. Werner-Allen, G., Lorincz, K., Ruiz, M., Marcillo, O., Johnson, J., Lees, J., Welsh, M.: Deploying a wireless sensor network on an active volcano. IEEE Internet Comput. **10**(2), 18–25 (2006)
91. Williams, W.: How friday's cyberattack shut down netflix, twitter, and spotify (2016). http://www.csmonitor.com/Technology/2016/1023/How-Friday-s-cyberattack-shut-down-Netflix-Twitter-and-Spotify
92. Wu, T., Disso, J.F.P., Jones, K., Campos, A.: Towards a scada forensics architecture. In: Proceedings of the 1st International Symposium on ICS & SCADA Cyber Security Research 2013, pp. 12–21. BCS (2013)
93. Xu, N.: A survey of sensor network applications. IEEE Commun. Mag. **40**(8), 102–114 (2002)
94. Yang, H., Zhuge, J., Liu, H., Liu, W.: A tool for volatile memory acquisition from android devices. In: IFIP International Conference on Digital Forensics, pp. 365–378. Springer (2016)
95. Zaharis, A., Martini, A.I., Perlepes, L., Stamoulis, G., Kikiras, P.: Live forensics framework for wireless sensor nodes using sandboxing. In: Proceedings of the 6th ACM Workshop on QoS and Security for Wireless and Mobile Networks, pp. 70–77. ACM (2010)
96. Zawoad, S., Hasan, R.: Faiot: Towards building a forensics aware eco system for the internet of things. In: 2015 IEEE International Conference on Services Computing (SCC), pp. 279–284. IEEE (2015)

Dependable Wireless Communication and Localization in the Internet of Things

Bernhard Großwindhager, Michael Rath, Mustafa S. Bakr, Philipp Greiner, Carlo Alberto Boano, Klaus Witrisal, Fabrizio Gentili, Jasmin Grosinger, Wolfgang Bösch and Kay Römer

Abstract Wireless technologies suffer from physical and man-made impairments (e.g., multipath propagation and interference from competing transmissions, as well as from the effect of temperature variations and other environmental properties): this impairs the reliability, timeliness, and availability of IoT systems. At the same time, we see a wave of new safety-critical IoT applications that require performance guarantees. This chapter surveys methods to increase the dependability of the IoT,

B. Großwindhager · C. A. Boano · K. Römer (✉)
Institute of Technical Informatics, Graz University of Technology,
Graz, Austria
e-mail: roemer@tugraz.at

B. Großwindhager
e-mail: grosswindhager@tugraz.at

C. A. Boano
e-mail: cboano@tugraz.at

M. Rath · K. Witrisal
Signal Processing and Speech Communication Laboratory, Graz University
of Technology, Graz, Austria
e-mail: mrath@tugraz.at

K. Witrisal
e-mail: witrisal@tugraz.at

M. S. Bakr · P. Greiner · F. Gentili · J. Grosinger · W. Bösch
Institute for Microwave and Photonic Engineering, Graz University of Technology,
Graz, Austria
e-mail: mustafa.bakr@tugraz.at

P. Greiner
e-mail: philipp.greiner@tugraz.at

F. Gentili
e-mail: fabrizio.gentili@tugraz.at

J. Grosinger
e-mail: jasmin.grosinger@tugraz.at

W. Bösch
e-mail: wbosch@tugraz.at

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_7

specifically focusing, first, on increasing the frequency bandwidth from narrowband, over wideband, towards ultra-wideband to better handle multipath effects and interference. Second, the chapter focuses on increasing the adaptability such that a networked system can compensate disturbances also dynamically, eventually striving for cognitive abilities. A distinguishing feature of this chapter is its comprehensive treatment of dependability issues across multiple layers, from signal processing, over microwave engineering, and to networking.

1 Introduction

Wireless technologies enabling the Internet of Things (IoT) are gaining momentum and pave the way for applications with high societal relevance and impact. Application domains include smart cars communicating to each other and with the road infrastructure for a safer drive, smart factories controlling and optimizing production processes, smart grids improving the efficiency of the distribution and consumption of energy, as well as smart buildings maximizing the comfort of its inhabitants while reducing the monthly energy bill. All the aforementioned applications impose vastly diverse requirements on system performance, ranging from highly efficient operation in order to maximize the lifetime of battery-powered devices, to ultra-low communication latency in order to enable high responsiveness. As an example, safety-critical IoT systems used to build smart cars, smart cities, or smart factories, require a high *reliability* and *timeliness*, as opposed to IoT systems for long-term monitoring that instead demand energy-efficient operations and hence a high *availability*. Reliability (continuity of correct and accurate service), availability (readiness for correct service) and timeliness (continuity of timely service) are key *dependability* attributes affecting the performance of an IoT system. Combined together with safety, confidentiality, and integrity, those attributes summarize the properties of a system and specify how much a user can rely on and trust its operations [1, 2].

Both research community and industry have long striven to produce solutions at all ISO/OSI layers to maximize these three key dependability metrics. At the physical layer, the focus has been on the design of highly efficient radio transceivers minimizing energy consumption and bit error rate, while maximizing range and throughput. This includes the design of reconfigurable filters as well as antennas operating on an increasingly broader frequency spectrum.

At the signal processing level, significant work has been carried out to increase the robustness of wireless communication and minimize the influence of multipath propagation and fading, as well as to achieve accurate indoor localization. The accuracy of the latter particularly affects the reliability of IoT applications that combine accurate positioning information with business logic. Examples of such applications include dynamic personalized pricing, product placement, and advertisement [3], the estimation of the popularity of exhibitions [4], as well as supply chain management and logistics.

At the networking level, a "soup" [5] of communication protocols have been developed in order to allow multiple devices sharing the same medium to communicate in a dependable fashion. These protocols address not only efficient operations allowing to increase the system lifetime and hence its availability, but also the ability to avoid collisions with other devices operating in the same frequencies, hence increasing the reliability and timeliness of communications.

All this work led to a plethora of wireless technologies and standards being proposed and commercialized in the last decade specifically for IoT devices. Technologies such as Bluetooth low energy (BLE), DigiMesh, and Z-Wave address IoT applications requiring low-range ultra-low-power communications. Newly developed standards such as Wi-Fi HaLow (IEEE 802.11ah), LoRA, SIGFOX, Weightless, On-Ramp Wireless, and NarrowBand IoT allows to form wide area networks on a large scale. WirelessHART, ISA100.11a, ANT+, and IEEE 802.15.4e, instead, focus on reliable and timely communications and hence specifically target safety-critical and industrial settings, whereas IEEE 802.15.4a explicitly supports ranging and is hence, well-suited for IoT applications requiring accurate localization.

In this chapter, the broad spectrum of solutions that have been developed in the past decades to achieve dependable wireless communication and localization in the Internet of Things is analyzed. Different from works that address the design of dependable IoT solutions only on individual layers, this chapter gives a comprehensive view that spans across different disciplines: from microwave engineering, over signal processing, to networking. This broader scope allows us to show that the efforts in increasing the bandwidth and the adaptability of the system (i.e., its ability to adapt and react dynamically to changes) do help in increasing the overall dependability of an IoT system. This concept is highlighted by outlining the design space of the solutions developed within the aforementioned disciplines. The discussions relate the adaptability of the solution (i.e., whether it is static, switchable, reconfigurable, reactive, predictive, or cognitive) to the employed bandwidth (i.e., narrowband, wideband, or ultra-wideband), and describe the different trade-offs from a high-level perspective. It is observed that dependability generally increases as the level of adaptability increases. For example, a networked systems that is aware of specific disturbances in the environment can adapt to counteract these disturbances. While a switchable or reconfigurable system can be adjusted by the operator to fit a certain environment, reactive systems do automatically perform such an adjustment based on past observations, and predictive systems employ sophisticated models to extend the time horizon across which disturbances can be anticipated. Ultimately, the aim is to strive for cognitive networked systems that perceive their state and environment and adjust its behavior in a smart way. The second dimension, bandwidth, is also crucial for dependability. By moving towards higher bandwidths, narrowband disturbances and signal propagation effects, such as multipath fading, can be better handled. The chapter shows that the ongoing efforts towards the design of cognitive solutions operating on ultra-wideband promise to increase the dependability of IoT communication and localization even further, although plenty of work is still required in this direction.

This chapter proceeds as follows. Section 2 provides a description of how the propagation of radio signals affects the robustness of a wireless link, and shows how these effects are related to two parameters of the design space, namely signal bandwidth and radio adaptability. Section 3 follows up with a description of the signal processing task for a single radio link, on the one hand, to enable reliable communication over the link, and on the other hand, to determine the position of the involved IoT devices. Section 4 provides an overview of modern transceivers which are used to build IoT applications and their most important building blocks, namely filters and antennas. Section 5 discusses the impact of medium access control layer protocols on key dependability attributes such as timeliness, availability, and reliability. It then introduces the design space in the networking domain, showing that there is still plenty of room to improve the dependability of IoT systems operating at high bandwidth. Finally, Sect. 6 wraps up our analysis and summarizes our conclusions.

2 Fundamentals of Wireless Propagation Channels

The reliability of a wireless link between two IoT nodes is mainly influenced by two factors: the physical radio channel (including multipath propagation and additive white Gaussian noise (AWGN)) and man-made interferences (from competing transmissions and unintended radiations). In this section, the focus is placed on the former, showing how multipath propagation affects the radio signal sent from a transmitter (TX) to a receiver (RX). In typical IoT application scenarios, it is intended to use low-cost and low-power devices which are faced with cluttered environments resulting in dense multipath that significantly deteriorates the transmitted signals. Hence, to develop dependable IoT systems, it is crucial to take the multipath propagation effects into account. Without modeling these channel imperfections and without taking them into account in the design process, it is not possible to give guarantees on the overall performance of the system. The discussion will lead to a system classification with respect to the signal bandwidth, demonstrating that this parameter has a major influence on the reliability of a wireless link. The need for adaptability is also highlighted, addressing temporal and spatial variations of the channel. The signal processing required to implement a reliable wireless link is discussed in Sect. 3.

Interfering radio signals will undergo similar propagation effects. To mitigate the interferences, a typical IoT radio will rely on the transceiver front-end (which is discussed in Sect. 4) and the MAC protocol (see Sect. 5). This is the most efficient way forward to achieve low-power operation and low complexity and thus availability.

2.1 Modeling of Multipath Channels

When one examines the radio signal seen by the receiver, it can be observed that a change of the receiver position relative to the transmitter results in fluctuations of

the received signal power. These fluctuations are referred to as fading [6, Chap. 2.1]. The covered distance between TX and RX in relation to the used carrier wavelength λ determines different *scales of fading*.

For a distance variation greater than approximately 10λ, the power variations are denoted as *large-scale fading*, which is characterized by two mechanisms, *path loss* (PL) and *shadowing*. PL indicates how much the received signal is attenuated with increased distance d between TX and RX. PL is described by an exponential decrease of power, which can be defined by a loss of $n \cdot 10\log(d)$ dB. The loss exponent n depends on the considered environment. For example, in free space, without any interacting objects, the loss exponent would be $n = 2$ described by Friis' equation [7, Chap. 3.9]. In line-of-sight (LOS) indoor scenarios, it can be below 2 due to constructive interference of reflections, while for environments with many obstacles, values between 2 and 6 have been observed [8, Chap. 3.6].

Shadowing is caused by obstacles in the environment blocking off signal paths and thus, leading to deviations from the deterministic PL law. It causes variations in the received power that are often described by a log-normal distribution, i.e., a Gaussian distribution in dB. For more details on large-scale models, the reader is referred to [6–8].

For small distance changes of less than approximately 10λ, the resulting received power fluctuations are called *small-scale fading* (*SSF*) which is due to multipath propagation. Multipath is caused by reflections and scattering of the transmitted signal within the physical environment. The receiver sees the signal from the direct path—the LOS component, overlapped with reflections called multipath components (MPC). These MPCs have different phases resulting in constructive or destructive interference, which is known as multipath fading. Mathematically, the multipath propagation can be described by the time-variant channel impulse response (CIR):

$$h(\tau; t) = \sum_{i=0}^{\infty} \alpha_i(t) e^{j\phi_i(t)} \delta(\tau - \tau_i(t)), \tag{1}$$

given in baseband equivalent form, where τ describes the delay time while t describes how the channel (the environment) changes over time. This formula describes the effect of (theoretically) infinitely many signal paths between TX and RX. The length of each path determines its time delay, denoted by the Dirac pulses $\delta(\tau - \tau_i(t))$. This delay also determines the phase shift $\phi_i(t) = -2\pi f_0 \tau_i(t)$, where f_0 is the carrier frequency, and the amplitude $\alpha_i(t)$ that decreases due to PL. Both amplitude and phase also experience variations due to interactions with the physical environment, i.e., reflections from different materials [8, Chap. 5].

The impact of the CIR (1) on radio signals can be analyzed by probing the channel at frequency $f_0 + f$, using the (baseband) signal $s(t) = e^{j2\pi ft}$. Convolution with (1) yields the received signal $r(t) = h(f, t)s(t)$, where $h(f, t)$ is a time–frequency-variant *channel gain* which describes the multipath interference experienced as SSF. The MPCs add up with different phases. Small movements of the RX change each phase differently, which causes rapid fluctuations of the channel gain as illustrated in

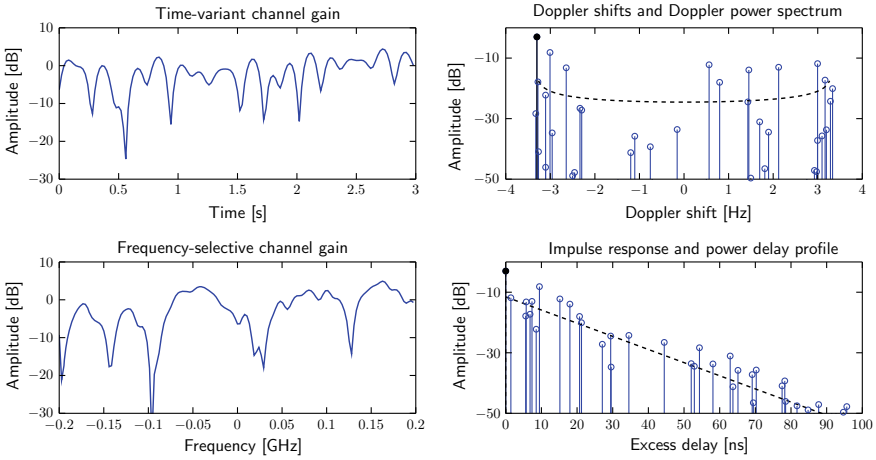


Fig. 1 Channel gain variations in time/frequency domain (left) and the resulting Doppler shifts/delay dispersion (right)

the top left of Fig. 1. The channel gain is similarly influenced by a shift in frequency, illustrated in the bottom left of Fig. 1.

To characterize the SSF, statistical models are used, justified by the large number of MPCs. For the assumption that the MPCs have uniformly distributed phases and there is no "dominant" MPC with exceedingly large amplitude, the SSF can be described with a *Rayleigh model*: The central limit theorem states that the channel gain will be a complex Gaussian random process because it is the summed effect of a large ensemble of MPCs [9]. Correspondingly, its amplitude distribution will be described by a Rayleigh probability density function [6]. It is needless to say that the amplitude (and power) gain of the radio channel is of key importance for the performance and thus the dependability of a wireless link between IoT nodes.

Time variations of the channel gain are (for example) caused by movements of the TX or RX. The upper left of Fig. 1 shows such variations caused by the RX moving at a constant velocity v. The Fourier transform of this time-variant gain is shown in the top right part of the figure. It can be seen that frequency shifts, so-called Doppler shifts, are introduced for each MPC i. Each Doppler shift is determined by $\nu_i = \frac{v}{\lambda} \sin(\theta_i)$, where θ_i describes the angle of arrival (AoA) of the i-th MPC. Hence, it is related to the speed of movement, the carrier frequency (which determines the wavelength $\lambda = c/f_0$), and the AoA. The dashed line shows the *Doppler power spectrum* (DPS) which quantifies the mean power output of the channel at a certain Doppler shift and thereby characterizes the time variations statistically.

An important parameter that can be derived from the DPS is the *coherence time* of the channel. It is indirectly proportional to the width of the DPS and determines how long the time frame is in which the channel gain is (widely) unaffected by Doppler effects. For system design, this sets an upper limit on packet duration before the TX/RX needs to readapt to a new channel gain. The illustrated example corresponds to a maximum Doppler shift of ± 3 Hz and a coherence time of about 0.5 seconds.

In the same fashion, one can regard variations of the channel gain in the frequency domain as shown in the bottom left of Fig. 1. Applying the Fourier transform results in the impulse response shown in the bottom right, the shown impulses are related to the MPC delays τ_i and amplitudes α_i which were described before. It should be noted that the term *excess delay* on the axis means that the time when the LOS arrives at the RX has been set to zero.

The dashed line indicates the statistical characterization of the impulse response, named the *power delay profile* (PDP). The PDP determines the *average* power output of the channel as a function of the delay time. The broadness of the PDP, and thus the *dispersiveness* of the channel indicates how long-lasting and powerful reflections and scattering in the environment are, i.e., it shows how much a transmitted waveform is spread out in time. This is called time dispersion.

The frequency variations are characterized by the *coherence bandwidth* related to the reciprocal of the PDP broadness. It indicates how fast the channel changes in the frequency domain due to multipath propagation. This parameter will be used later to distinguish narrowband and wideband systems, which have flat-fading and frequency-selective channels, respectively. The shown example relates to a coherence bandwidth of about 10 MHz.

Spatial variations can be regarded similarly since a movement of the TX or RX also leads to phase variations of the MPCs and, therefore, to multipath fading. The amount of the variations is related to the wavelength and the distribution of the AoAs. The *coherence distance* of the channel is a parameter which is important for designing antenna placements in multi-antenna systems as will be shown later.

2.2 Signal Classification in Terms of Bandwidth

The impact of the propagation channel on a transmitted signal can be classified with respect to the signal bandwidth. The propagation of a transmitted signal $s(t)$ is described as a linear filtering with the time-variant channel impulse response $h(\tau; t)$ or—in frequency domain—as a multiplication with the channel gain $h(f; t)$. The signal is modulated onto the carrier f_0 and B denotes the biggest deviation in the band, thus the signal bandwidth ranges from $f_0 - \frac{B}{2}$ to $f_0 + \frac{B}{2}$.

When a *narrow bandwidth* (*NB*) is used for the transmitted waveform with a bandwidth less than the coherence bandwidth, the channel response is flat over the respective band. More specifically one has $h(f; t) \approx h(t)$ for $f \in \left[f_0 - \frac{B}{2}, f_0 + \frac{B}{2}\right]$, which means the only effect of the channel is an attenuation of the original signal by $h(t)$. This is called *flat fading*. However, the value of $h(t)$ fluctuates due to the Rayleigh fading as described before. Figure 2 indicates the flat-fading case on the left-hand side, with a nearly constant spectrum in the (small) region between the dashed lines.

When a signal with a significantly *wider bandwidth* (*WB*) is used, a filtering (multiplication of spectra) results in a larger part of the channel frequency response being "visible" or "cut out". The resulting spectrum is not flat anymore, which is

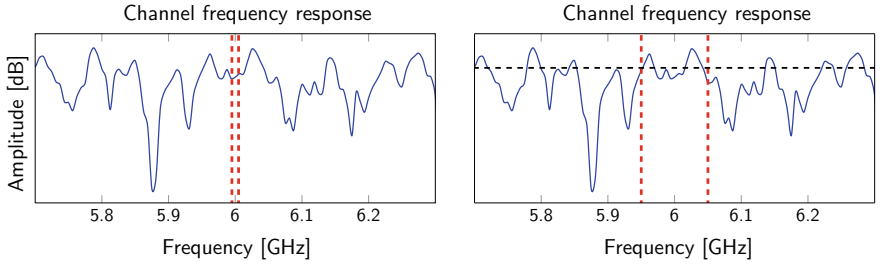


Fig. 2 Effect of propagation channel on NB ($B = 10$ MHz) and WB ($B = 100$ MHz) pulses in the frequency domain. Vertical dashed lines show the frequency range of the transmitted signal

known as a *frequency-selective* channel. An example is shown in the right-hand side of Fig. 2. In time domain, the effect corresponds to linear distortions which will introduce, for example, intersymbol interference (ISI) that has to be dealt with in receiver systems. In terms of system design, the coherence bandwidth defines the limit between NB and WB.

For even higher bandwidths, the region of *ultra-wideband* (*UWB*) is reached. Standardization bodies have introduced definitions of *absolute UWB* for bandwidths of more than 500 MHz and *relative UWB* for bandwidths larger than 20% of the carrier frequency [10, 11]. These definitions are not clearly related to physical properties of the radio channel in contrast to the boundary between NB and WB. However, the average power over a bandwidth in this range is almost constant (indicated by the dashed line in the right-hand side of Fig. 2), regardless of the channel realization, i.e., the small-scale variations of the power gain become negligible, thus a dependable link can be provided.

The UWB boundary is also related to time resolution properties. One can speak of UWB when MPCs can be resolved in time domain, e.g., when the LOS component does not overlap with any MPCs arriving later. This is illustrated in Fig. 3, where the effect of the channel onto a WB and a UWB pulse is shown in time domain. The red, dashed curve shows the transmitted pulse, which indicates its length in relation to the CIR. It can be seen that in the WB case, resolvable time taps equivalent to the pulse width still contain many MPCs and thus they are strongly influenced by multipath fading.

3 Physical-Layer Signal Processing

Following up from the discussion on the wireless channel in Sect. 2, the main *signal processing* task in an IoT radio concerns overcoming the impact of physical propagation effects, in particular, multipath fading. Compared to traditional wireless systems, IoT systems may operate in cluttered environments with lots of obstacles. This

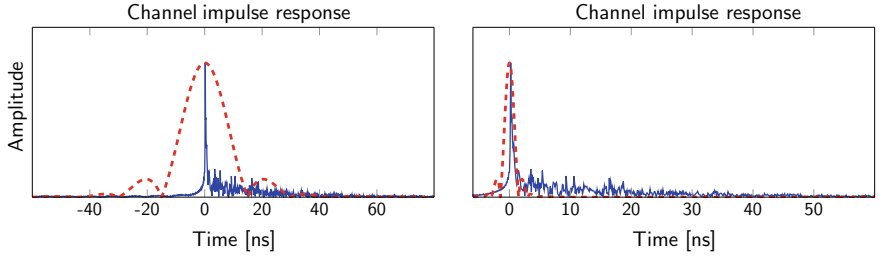


Fig. 3 Effect of propagation channel on WB ($B = 100$ MHz) and UWB ($B = 1$ GHz) pulses in the time domain

produces a high amount of multipath components which could interfere destructively with the line-of-sight component and possibly results in weak links.

Two related receiver signal processing tasks will be addressed, namely the recovery of the transmitted data in multipath-rich environments (see Sect. 3.1) and the measurement acquisition for positioning (Sect. 3.2). For each, the design space will be outlined, showing how the bandwidth and adaptability influence the reliability and availability of the wireless links and position estimation. It will be argued that the adaptability is related to the channel knowledge of the radio nodes, which is used to adapt the RX and/or the TX to the current channel state.

3.1 Signal Processing for Wireless Communications

To address the data recovery, it will be shown how communication systems can be designed to deal with fading effects. Figure 4 presents the design space, classifying physical-layer processing schemes regarding their bandwidth and adaptability. The bandwidth is represented by the classification given in Sect. 2.2, while the adaptability is related to the channel knowledge that must be available at the RX and/or the TX in order to obtain a reliable radio link. A cognitive radio also takes interference of competing users into account, which is denoted as "environment" knowledge. This box (and the "time-reversal" scheme) are shown in gray because these approaches are considered to be "less practical", in particular, under the complexity constraints of an IoT. An increase of bandwidth and/or adaptability helps in general to overcome the fading effect, mitigate interferences and thus leads to an improvement of the reliability of wireless radios.

Narrowband transmission will be predominant in many IoT applications, due to their low complexity. In fact, a "fixed" RX architecture can be used in this case, shown in the bottom left corner of the design space. The optimum receiver design for this case would be a *matched filter* (MF) with impulse response $h(-t) = s(t)$ [12, Chap. 7.5], [13, Chap. 5.3], matched to the transmitted waveform denoted by $s(t)$ which is the only thing that must be known at the receiver. The sampled output

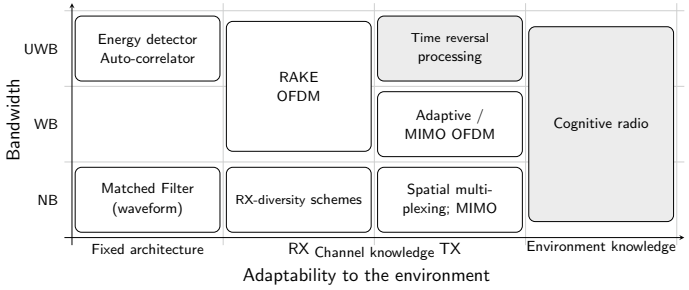


Fig. 4 Design space for systems to mitigate fading in wireless propagation channels

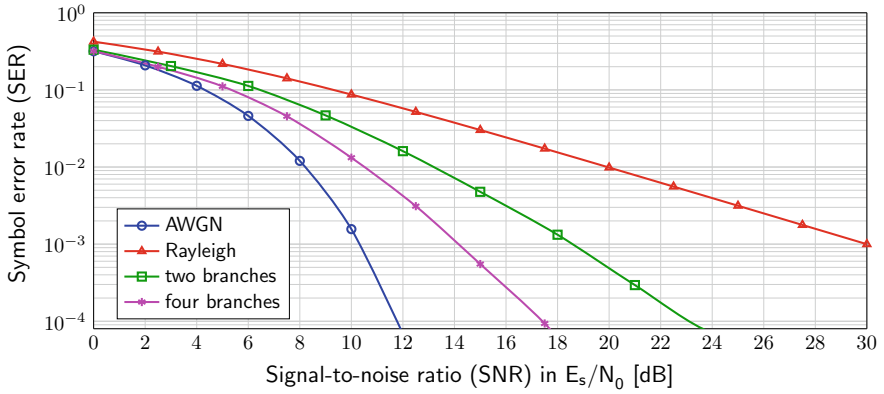


Fig. 5 Comparison of AWGN with Rayleigh channel for single and multiple antennas

of the MF can be modeled by:

$$r = \sqrt{E_s}h \cdot s + n, \tag{2}$$

with h describing the complex-valued gain of the flat-fading (NB) channel, the energy E_s of the transmitted symbol s, and a noise sample n. Due to the multipath fading, the gain $|h|$ exhibits a Rayleigh distribution. Figure 5 depicts the required signal-to-noise ratio (SNR) to reach certain average symbol error rates (red curve) in comparison to a nonfading (h is constant) AWGN channel performance.[1] It can be seen that there is a large gap. For example, to achieve an error rate of 10^{-3}, one needs approximately 20 dB more SNR for the (Rayleigh) fading channel. Hence, the multipath fading has a strong impact on the availability and reliability of wireless links and thus its treatment is also highly relevant for designing dependable IoT applications.

[1]The AWGN is characterized by a flat double-sided power spectral density of $N_0/2$.

3.1.1 Overcoming the Fading Effect: Diversity Principle

This subsection considers how to overcome the previously described performance gap. The fading is addressed by *diversity schemes*, which make use of multiple observations of the TX signal over independent paths (or branches) with channel gains $h_i, i = 1 \ldots M$. When the channel gains of each branch are known at the RX, the observations can be processed according to the *maximum ratio combining* (MRC) principle [14, Chap. 9.1], which can be seen as "correcting" the received branches for their suffered phase shifts through the Rayleigh channel, allowing to add the branch contributions coherently. Furthermore, proper weighting is applied to maximize the resulting SNR (diversity gain). Making use of independent channel gains is essential for exploiting diversity. It results in a significant reduction of the performance gap as illustrated in Fig. 5. Independent channels can be found in various domains, as outlined in the following.

Time/Frequency diversity. Depending on the coherence time and bandwidth, different observations spaced in time or (carrier) frequency can be combined. However, this means that the same signal is sent multiple times, thus, assuming one has the energy E_s available, the SNR effectively decreases by $1/M$.

Antenna (RX) diversity. In the spatial domain, multiple receiving antennas (spaced farther than the coherence distance) can be used to combine their measurements and obtain the diversity gain. The advantage here is that the SNR is not decreased; a single transmission is sufficient. However, more hardware and thus complexity is needed with M_R RX antennas.

TX diversity. At the TX side, using M_T transmit antennas, one has to prepare the transmitted signals to be able to separate the individual branches at the RX. There are two different methods:

- Pre-coding: Using encoding and transmission sequences, the branches can be separated even without channel knowledge available at the TX [15].
- Pre-filtering: This method is equivalent to the RX diversity scheme, the ratios from the MRC scheme are applied before transmission, hence, channel knowledge is required at the TX.

Systems that use multiple TX and RX antennas are called *Multiple-Input Multiple-Output* (*MIMO*) systems. Concerning diversity, a MIMO system is capable of achieving a diversity gain related to the product $M_T \cdot M_R$. However, it can also multiplex $(\min[M_T, M_R])$ parallel data streams to the same frequency band at a given time. A prerequisite for MIMO signal processing is channel knowledge at the TX and RX. The more channel knowledge available, the higher the spectral efficiency and reliability of the radio link. However, the required high adaptability to the channel at the TX and RX comes at the cost of higher complexity. The application of (adaptive) directive antennas can yield—in principle—similar gains. Directive antennas could thus be a low-complexity alternative to multi-antenna systems for the IoT.

3.1.2 Wideband Systems

Instead of introducing multiple antennas, one can also exploit frequency diversity from a WB signal to increase the reliability. In the WB case, the effect of the channel is described by a linear convolution with the CIR, leading to linear distortions of the transmitted signal, ISI, and (still) multipath fading. For these signals, channel knowledge is always required to *efficiently* deal with the channel effects.

The received signal is spread over different delay bins because of the delay dispersion. A bank of correlators which is called *rake receiver* [6, Chap. 18.2.4] can be used to collect the energy at different delay bins (rake "fingers"). At each finger, the received signal energy will still suffer from multipath fading because each finger sees the combined effect of a large number of multipath components as shown in Fig. 3 (left-hand side). However, the rake receiver can combine these delay branches using MRC. It thereby exploits frequency diversity and can thus decrease the performance gap of the symbol error rate. The rake receiver is usually employed with spread spectrum signals which occupy more than the minimum Nyquist bandwidth, given some desired data rate. Spread spectrum systems are also robust with respect to NB interference, hence improving the dependability [16].

Orthogonal frequency-division multiplexing (*OFDM*) is a multicarrier technique where the transmitted WB signal is split into NB subchannels, which are densely packed in frequency, exploiting orthogonality properties [17]. It is the method of choice for high-rate transmission at high spectral efficiency, avoiding the ISI, fading, and signal distortion issues in an elegant way. Again, frequency diversity is exploited to increase the robustness to frequency-selective fading, using coding and interleaving to correct for deep fades affecting certain subchannels. OFDM is used in state-of-the-art high data-rate communication standards, e.g., in wireless LANs according to IEEE 802.11a, g, n, ac [18] and in LTE [19, Chap. 3].

When channel knowledge is also available at the TX, an adaptive ("waterfilling") OFDM technique can be used [20] where the modulation order of the data symbols is adapted to the SNR at individual subchannels. OFDM is also the method of choice to implement MIMO systems in frequency-selective (WB) channels [21].

3.1.3 Ultra-Wideband Systems

Moving to the top row in the design space, sufficient bandwidth is available to achieve the time resolution to resolve individual MPCs and mitigate multipath fading to a large extent. For communications, the same principles can be used as in the WB case (rake, OFDM). The available frequency diversity will be high enough to reach a performance close to the AWGN channel, which is a key characteristic of a UWB system (see Sect. 2.2 and [22, 23]). Hence, it is not needed to implement multiple antennas in order to overcome the multipath fading. However, the practical hardware implementation is still problematic. The RX processing requires high sampling rates leading to high power consumption (see Sect. 4.1).

Systems that tackle the high-cost problem are situated in the top left corner of the design space: noncoherent UWB receivers with simple structures like the *energy detector* (*ED*) or the *autocorrelation receiver* (*AcR*). The basic function of these systems is to multiply the received signal with itself (or a delayed version, in case of the AcR) followed by an integration over a certain time frame. Via the multiplication, the carrier phase information is lost and only the envelope of the signal is obtained. The integration yields an accumulation of energy for a specific time window. Hence, transmitted symbols can be detected in certain time frames, enabling time-hopping schemes for multiple access [24, 25]. In case of the AcR, the delay is an additional tuning parameter which allows for more sophisticated transmission schemes such as transmitted-reference [26, 27] where time-hopping pulse sequences can be detected. The ED method is also supported by the IEEE 802.15.4a standard (see Sect. 5.2) which describes a UWB air interface for sensor networks and (indoor) positioning.

Moving towards the top right of the design space means that on top of high bandwidths, more channel knowledge is available. This allows the application of the conceptual method of *time reversal* (*TR*). It is assumed that the TX knows the CIR, which is then time reversed and is used as a prefilter to transmit a pulse. This results in all MPCs arriving at the same time instant and adding up in a constructive way, resulting in a high SNR with simple processing at the RX [28].

Finally, the concept of *cognitive radio* should be mentioned. Radio nodes are assumed to be aware of their surroundings (in this case, the radio environment) and adapt their state to it. In addition to channel state information, cognitive radio also takes interferences of other radios into account, which again increases the amount of available information and also the dependability of the wireless links [29–31].

3.2 Signal Processing for Wireless Localization

Localization (or positioning) describes the process of estimating the position of mobile devices in a defined coordinate system. Radio signals transmitted between "agents" and "anchors" can be used for positioning, in environments where satellite-based systems are useless, for example, indoors. Such radio positioning systems "measure" parameters of the received signal which are related to the geometry of the arrangement of the radio nodes, for example, the time of flight (ToF) or the received signal strength (RSS) which relates to the distance, or the angle of arrival (AoA).

It remains a tremendous challenge to obtain a dependable (indoor) positioning system, one that has sufficient accuracy and reliability so that for example, the navigation of an automated vehicle could rely upon it. Multipath propagation is a key reason hindering the implementation of an accurate and robust positioning system. In this section, the impact of multipath propagation on the measurement acquisition and positioning tasks is discussed.

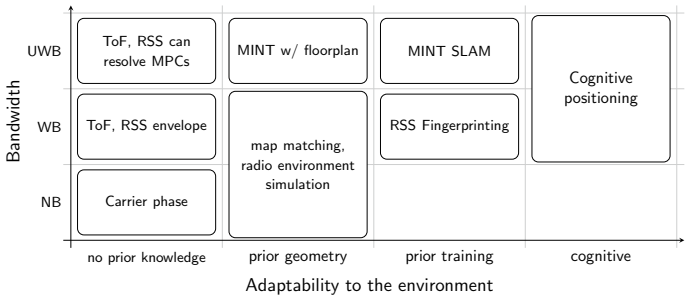


Fig. 6 Design space for dependable positioning systems

3.2.1 Overview of Dependable Positioning Systems

A taxonomy of positioning approaches is found in Fig. 6 in a design space that again spans the signal/system bandwidth and the adaptability of the system. The great advantage of a large bandwidth (UWB) system lies in the fact that the line-of-sight (LOS) component of the channel response can be separated from the multipath, hence its parameters, in particular, the ToF and the free-space path loss can be measured accurately. At a lower bandwidth, many MPCs will interfere with the LOS, which will introduce fading and pulse distortions and thus reduce the potential accuracy. These properties of the radio channel have already been introduced in Sect. 2.2. An in-depth analysis of the influence of bandwidth on ToF ranging will be given below.

With an RSS approach, one can for instance use *fingerprinting*, an approach rooted in machine learning. In the first step, a database is created that maps certain positions to RSS values from multiple anchors. This map then allows wireless nodes to associate any measured RSS values with a certain position [32]. The system is configured for a certain environment. Map matching is an approach, where prior information about a building floor plan is used to avoid position fixes that disagree with the geometric constraints of an environment. *Multipath-assisted indoor navigation and tracking* (MINT) also uses floor plans. More specifically, it associates reflected MPCs with the environment requiring UWB signals to achieve sufficient time resolution [33]. A simultaneous localization and mapping (SLAM) approach learns a suitable feature map online, exploiting past measurements of the environment, which is then used for the self-localization [34–36]. More prior information, in general, enhances the performance and thus supports the goal of dependable positioning.

A cognitive positioning system, finally, also learns environment information on its own [31, 37]. It goes beyond the capabilities of a SLAM algorithm in that it implements "cognitive" features such as active feedback on the environment, attention, and memory. It can for instance, schedule measurements in a way that the expected information gain is maximized, it can focus on relevant information—consider a system that has to deal with an abundance of clutter measurements—and it can use a hierarchically structured memory to allow for different layers of abstractions. This

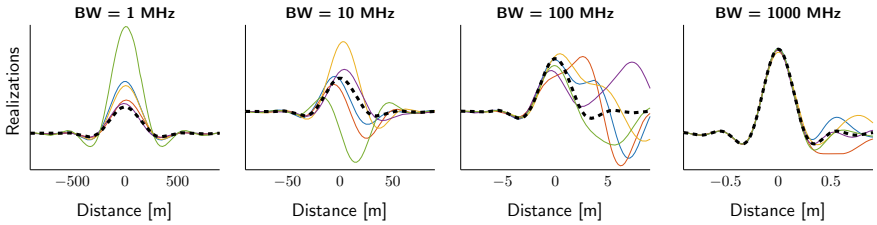


Fig. 7 Correlator outputs for different bandwidths (BW) and positions (realizations of the channel)

approach is found to the very right of the taxonomy. It is a research topic that is widely unexplored as of today, in particular, under the low-power constraints of the IoT.

3.2.2 ToF Positioning in a Multipath Environment

ToF methods are robust in the sense that the accuracy depends only little on the actual range between transmitter and receiver, as long as the SNR of the received signal is sufficient. In the following, the impact of multipath on such ToF-based positioning systems is explored. While the focus lies on ToF, some of the conclusions will generalize to other measurement methods as well.

The bandwidth of an RF signal determines its time resolution. An optimal estimator for the ToF in AWGN simply correlates the RX signal with the transmitted signal that is assumed to be known for this purpose (consider, e.g., a training sequence). The duration of the main lobe of the correlation function is directly linked to the time resolution.

Figure 7 shows examples of correlator peaks for different bandwidths and different channel realizations. The differences in the waveforms are due to variations of the multipath. It should be noted that the time-scales—encoded in terms of distance in meters—vary heavily according to the signal bandwidth. For comparison, the LOS without multipath is shown by the thick, dashed curves.

In the NB case, it can be seen that the transmitted signal has a length of hundreds of meters, which means that the LOS component overlaps with all later arriving multipath components, resulting in a flat-fading case as discussed before. Even though there is no pulse distortion here, the rise time of the TX signal is extremely long, spanning hundreds of meters or more, thus an NB signal is not suitable for accurate positioning. One could try exploiting the carrier phase for ranging instead of the NB envelope. This would tremendously improve the accuracy, but the big issues with that approach is the need for synchronized oscillators and calibrated radios, ambiguity among successive cycles of the carrier, and the random phase shift of a Rayleigh channel.

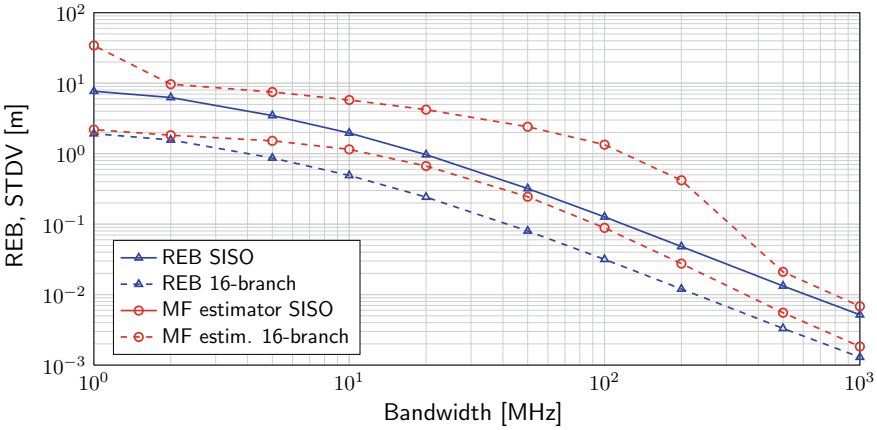


Fig. 8 Standard deviation (STDV) of the ranging error bound (REB) in relation to bandwidth for different diversity branch setups

Considering a bandwidth increase towards WB, the dispersiveness of the channel becomes visible which unfortunately distorts the pulse. Depending on the channel realization, this can significantly shift the correlation peak.

Going towards UWB yields an extremely narrow pulse and hence resolves the LOS component from other MPCs as described before. This would allow for an accurate measurement of the arrival time corresponding to cm-level accuracy.

A mathematical quantification of the BW dependence yields deeper insight. The Cramer–Rao lower bound (CRLB) [38] defines the theoretical lower limit on the variance of the estimated arrival time of the LOS component. In a dense multipath channel, the CRLB can be written as [39]:

$$\mathrm{var}\{\hat{\tau}\} \geq \left(8\pi^2\beta^2\widetilde{\mathrm{SINR}}\right)^{-1},$$

where β^2 is the mean-squared bandwidth of the signal and SINR is a signal-to-interference-and-noise ratio that accounts for the power of the interfering multipath and is thus *limited by the physical propagation environment*. The error variance scales reciprocally with the BW β^2. Furthermore, the interference from multipath also scales reciprocally with bandwidth, which additionally affects the ranging performance. An illustration of this result is given in Fig. 8. Between 10 MHz and 1 GHz, the *Ranging Error Bound* (*REB*) (consider the "SISO" curve) scales by more than a decade when changing the bandwidth by a factor of ten. At very low bandwidth the slope reduces [39], but here the accuracy is not at a useful level any more.

According to the REB, a bandwidth of 100 MHz is needed to obtain an accuracy in the 10-cm region. However, a simple matched-filter (MF) estimator of the ToF (which computes the cross-correlation with a known training sequence) cannot achieve this

bound as also illustrated in the figure. The MF will produce many positively biased outliers that cause a deviation from the theoretical bound.

Again, diversity can be exploited to improve the performance, assuming that multiple measurements are obtained over independent channel branches. It has been discussed in [39] that the number of independent branches multiplies—and hence increases—the available SINR. That is, a lack of bandwidth can be compensated to some extent by a diversity scheme. Note from Fig. 8 that also the simple MF estimator performs now closer to the theoretical bound. Indeed a standard deviation below 10 cm can now be achieved at a bandwidth of 100 MHz.

It is concluded that a higher bandwidth in particular, but also diversity from multiple measurements improves the reliability of a ToF estimator. Another advantage of the diversity approach lies in the availability that is improved by considering a larger set of independent measurements. While a single UWB link may provide the same accuracy, the risk of a severe outlier would be much higher, e.g., due to an obstruction of the LOS. The availability can be tackled with so-called *multipath-assisted methods* [40]. The solution is to "turn the enemy into an ally"; the enemy is multipath propagation that usually acts as a disturbance for ToF positioning. Multipath-assisted methods make use of detectable MPCs and treat them as separate observations originating from *virtual anchors* (VAs) [40]. These VAs can be determined when, e.g., the floor plan is known, moving towards the top right of the design space. Each surface that causes a clear specular reflection (such as a wall) yields a mirror image of the actual anchor which can be represented by the position of a VA. Hence, a single anchor "creates" a multitude of virtual sources as long as the pulses are short enough to be distinguishable. For longer pulses where less bandwidth is used, it was also shown that directional antennas could be used at the anchor to additionally explore the angle domain [41].

4 Hardware

This section elaborates on the hardware components that are an integral part of wireless IoT nodes. The section presents different transceiver structures that are used in wireless nodes and, in particular, their respective radio frequency (RF) stages and some of their most important components, i.e., filters and antennas that considerably mitigate interfering signals and thus enhance the IoT nodes' dependability. The presented discussions focus on these hardware components to realize dependability and in particular, on the components' impact on key dependability attributes such as availability, reliability, and timeliness. In Sect. 4.1, starting with an overview of modern transceivers used today in wireless IoT nodes, different transceiver structures and their specific RF stages are discussed based on the hardware design space, which is shown in Fig. 9. This discussion leads to the conclusion that dependable wireless nodes in IoT systems should ultimately rely on software-defined radio (SDR)-based or rather cognitive radio (CR)-based transceivers that are able to sense the whole frequency spectrum and adapt the wireless communication scheme in real time, thus

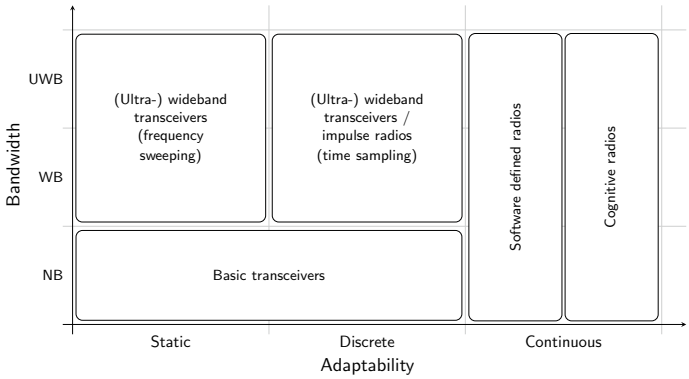


Fig. 9 Hardware design space for the transceivers of wireless IoT nodes

providing a high reliability and availability. To realize such transceivers, their filters and antennas also have to provide a continuous adaptability for the operation over a specific frequency range. These aspects of filters and antennas are discussed in Sects. 4.2 and 4.3, respectively, leading to the description of advanced filters and antennas for wireless IoT nodes, which enable different degrees of dependability for wireless IoT nodes.

4.1 Transceivers

The transceiver is a major component of wireless IoT nodes and thus, is a critical component to ensure dependable wireless communication and localization in IoT systems. Figure 10 shows a block diagram of a basic modern transceiver, i.e., the most commonly used transceiver configuration in wireless nodes nowadays [42]. The data of the node that is transmitted wirelessly to adjacent nodes in the network is generated and modulated using a dedicated digital signal processing hardware. The digital-to-analog converter (DAC) then converts the digitally modulated signal to the analog domain. The signal is then up-converted to the respective operating RF frequency of the wireless node by the RF transmitter and radiated by the antenna. Typically, one common antenna is shared by the RF stages of the transceiver, i.e., the RF transmitter and the RF receiver; an RF switch is used to provide decoupling between the transmitted and received signals. The received signal at the antenna of the wireless node is down-converted to baseband by the RF receiver, converted to the digital domain using an analog-to-digital converter (ADC), and demodulated and processed using dedicated digital signal processing hardware. In the following, the RF stages of the transceiver are discussed in detail, while the aspects of digital signal processing are discussed in Sect. 3.

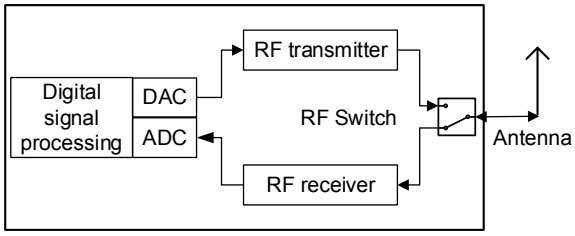
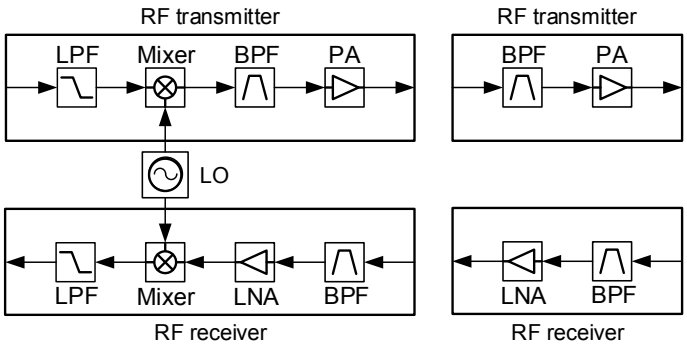


Fig. 10 Block diagram of a basic modern transceiver in wireless nodes

RF transmitter and RF receiver. The RF transceiver stages of most wireless systems have a high degree of commonality, even though there are many variations in practice [43]. Figure 11a shows the detailed block diagram of the RF stages of a basic modern transceiver typically used in IoT nodes. At the RF transmitter, the digitally modulated baseband signal is filtered by a low pass filter (LPF) and shifted up in frequency, i.e., the signal is up-converted to the desired RF operating frequency (e.g., to 2.45 GHz, see Sect. 5.2), using a mixer and a local oscillator (LO). A bandpass filter (BPF) allows the desired operating frequency to pass, while rejecting undesired frequencies generated during the upconversion. Subsequently, a power amplifier (PA) is used to provide the required transmitter output power, defined by wireless communication standards and regulations. Finally, the antenna converts the modulated carrier signal from the transmitter to a propagating electromagnetic wave to communicate wirelessly with an adjacent wireless node. The RF receiver retrieves the data transmitted by the adjacent wireless node, essentially reversing the functions of the RF transmitter components. The antenna receives electromagnetic waves radiated from many wireless nodes over a relatively wide frequency range. An input BPF provides some selectivity by filtering out received signals at undesired frequencies



(a) RF stages of a basic modern transceiver (b) RF stages of a SDR

Fig. 11 Block diagram of the RF stages of a basic modern transceiver and for the common implementation of the SDR concept as shown in [42]. See Fig. 10 for whole transceiver structure

and passing signals within the desired RF frequency band. The BPF is followed by a low-noise amplifier (LNA) that amplifies the possibly very weak received signal, while minimizing the noise power that is added by the amplification. Also, by placing the BPF before the LNA, the possibility is reduced that the sensitive amplifier will be overloaded by interfering signals of high power, generated, for example, by colocated wireless nodes that would impair the reliability of the IoT system. Next, the received RF signal is down-converted to the baseband by a mixer and the LO and filtered by an LPF.

4.1.1 Transceiver Adaptability

In this section, different transceiver structures for IoT nodes are investigated with respect to their adaptability, following the hardware design space introduced in Fig. 9. In particular, the investigation highlights two key attributes with respect to the dependability of wireless nodes, i.e., their availability and their reliability. In the following, the transceiver adaptability is defined by the adaptability of the RF operating frequency of the transceiver. The transceiver structure is static and provides no adaptability, if the transceiver operates at a fixed operating frequency or at multiple fixed operating frequencies. The transceiver provides a discrete adaptability, if the transceiver is able to switch between a fixed set of operating frequencies, to support, for example, blind and adaptive channel hopping network protocols and thus to enhance reliability. The transceiver provides a continuous adaptability, if the transceiver is able to operate at any frequency within a certain range that supports, for example, adaptive and predictive channel hopping protocols and high reliability (see Sect. 5.3.5).

With respect to a static transceiver structure, RF transmitter and receiver structures, as shown in Fig. 11a, are used that operate at a fixed operating frequency. Transceivers providing a multi-frequency operation can be realized by the design of an array of such RF transmitters and RF receivers, each operating on a specific frequency. Their combinations allows to operate at multiple fixed operating frequencies [44], thus providing no adaptability.

Transceivers for multi-frequency operation can also be designed using a single RF transmitter and receiver with a swept LO[2] [44], thus achieving a discrete adaptability. The RF stages shown in Fig. 11a can provide such a discrete adaptability using a swept LO. Using these RF stages with a swept LO, makes it necessary to design and realize BPFs, a PA, an antenna, as well as an LNA that supports the operation at the respective operating frequencies.

Continuous transceiver adaptability is provided by so-called SDRs. Currently, a lot of research effort is devoted to realize SDRs [42, 44–52]. The concept of an SDR has been first introduced by Mitola in 1995 [29], who proposed to create a radio that is fully adaptable by software in terms of operating frequency, bandwidth, and

[2]The RF operating frequency of the transceiver is switched between a fixed set of frequency values using a swept LO.

communication standard. A block diagram of the RF stages for a common implementation of the SDR is shown in Fig. 11b. The baseband signals are generated and up-converted by dedicated signal processing hardware, converted into analog waveforms by a DAC, filtered by a BPF, and amplified by a PA, before passing through an RF switch to be radiated by the antenna. The signal received by the antenna is routed to an LNA through the RF switch and a BPF and is then digitized by an ADC. Down-conversion, demodulation, and decoding are accomplished by dedicated signal processing hardware. To realize an SDR, a high sampling rate DAC has to be used at the transmitter side of the transceiver, as presented in [44]. At the receiver side of the transceiver, a so-called bandpass sampling receiver [42] has to be used exploiting a high sampling rate ADC.

Next to the high reliability given by the continuous adaptability of an SDR, another benefit of SDRs in IoT nodes are the possible hardware cost savings [42]. Currently, the drawback of SDRs in IoT nodes is their negative impact on the availability of the wireless nodes due to the use of power-consuming components (e.g., high sampling rate ADC) [42], considerably reducing the battery lifetime.

4.1.2 Transceiver Bandwidth

In general, the transceiver bandwidth is defined by the RF operating frequency band, which can be NB, WB, or UWB (see Sect. 2.2). Following the hardware design space in Fig. 9, this section presents two general transceiver concepts that provide a wide or rather ultra-wide transceiver bandwidth and thus may pave the way to dependable IoT systems, highlighting the IoT nodes' availability and reliability.

One UWB transceiver concept relies on transceiver structures that perform frequency sweeping. In comparison to the second concept presented below, a frequency sweeping-based UWB transceiver is currently the preferred implementation due to its more practicable architecture (i.e., balanced complexity of the RF transmitter and receiver structures). The transceivers divide the spectrum into several frequency bands, using, for example, a stepped-frequency continuous wave (SFCW) transmitter [53]. A SFCW-based transceiver transmits a series of discrete tones in a step-wise fashion to attain a large effective bandwidth. In frequency sweeping- based UWB transceivers, so-called sweeping receivers are used [54], exploiting a swept LO and thus a discrete transceiver adaptability. This kind of receiver uses electronically reconfigurable RF components to receive a large bandwidth by continuously sensing smaller portions of it [48, 53]. Examples of key reconfigurable RF components for these receiver architectures are reconfigurable BPFs for dynamic signal-band selections and reconfigurable notch filters for interference mitigation [47] as presented in Sect. 4.2. A major challenge within this type of receiver is the loss of phase information due to the sweeping operation [54]. To preserve this information and reconstruct the time-domain wideband waveform, an accurate calibration and special hardware design is required that is rather expensive [54], being a major drawback with respect to the low-cost requirements commonplace in the IoT.

Another UWB transceiver concept relies on transceiver structures that perform time sampling, also known as impulse radios. Time sampling-based UWB transceivers transmit a sequence of single, very short pulses at the whole bandwidth instantaneously [53]. This kind of UWB transceiver features a simple transmitter architecture with a low-power consumption. However, receiving the short duration UWB signals presents a considerable challenge. In these kind of UWB transceivers, sampling receivers are used, which capture instantaneously all the frequency components of the waveform by taking samples over time [54]. Within such receivers, the required high ADC sampling rate for signals occupying UWB frequencies often rules this approach out due to their high power consumption. This issue can be partially circumvented by means of properly conceived receiver architectures such as mixed-mode wideband receiver architectures, which hybridize the analog and digital domains [47] (e.g., the ED and AcR discussed in Sect. 3.1). The main advantage of sampling receivers is that the whole spectrum of the incident waveform is sampled at once [54]. Among many other things, this is important to acquire impulse response measurements as discussed in Sect. 3.2.

4.1.3 SDR/CR-Based Transceivers

Ultimately, future transceiver developments for wireless IoT nodes will evolve towards UWB transceivers that provide a continuous adaptability, realizing SDRs or rather CRs. A CR combines all the features of an SDR and adds more intelligence by sensing the radio environment and by tracking and adapting to changes in the wireless IoT system in real time [44]. The concept of CRs has also been introduced by Mitola [55], who states that SDRs provide an ideal platform for the realization of CRs. This concept has driven many researchers to study CR approaches [45–49]. Following the vision of SDR/CR-based transceiver in IoT nodes, most of the components of the RF transmitter and RF receiver stages will be shifted towards the digital domain (cf. Fig. 11a, b). However, some important components of the transceiver stages have to be still implemented in the analog domain as, for example, the BPFs, the PA, the antenna, the LNA, as well as the DAC and the ADC of the transceiver. This means that a lot of effort has to be put into the design of these components in order to provide highly reliable, highly available, and low-cost wireless IoT nodes. In the following subsections, filter and antenna realizations are investigated in more detail, which will be crucial to realize dependable wireless communication and localization in the IoT.

4.2 RF and Microwave Filters for the Internet of Things

As presented in Sect. 4.1, wireless transceivers rely on microwave and RF filters for different functions. Figure 11 shows that both RF transmitters and receivers make use of two types of filters: lowpass- (LPF) and bandpass filter (BPF). The reader is

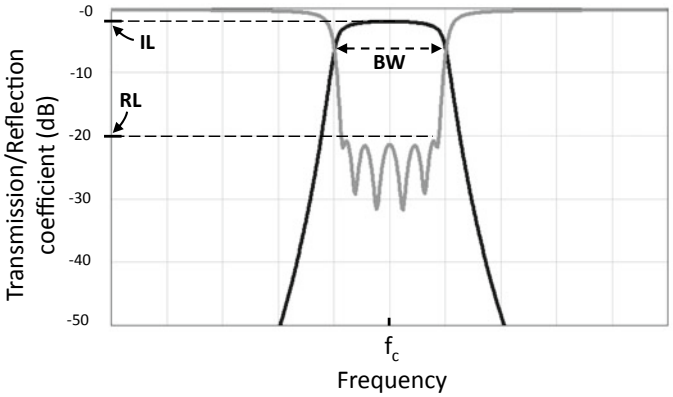


Fig. 12 Transmission and reflection coefficients of a 6^{th}-order BPF showing the main parameters to characterize a filter response

reminded that LPFs are circuits that block high frequencies and pass low frequencies, whilst BPFs let through only signals within a certain frequency band. Since the latter has a stronger link to dependability, the present section will be focused on BPFs, presenting their general aspects and potentials in the IoT domain. Challenges regarding their synthesis, physical realization, performance and tunability will be described in Sects. 4.2.1 and 4.2.2.

The main parameters used to describe the response of a BPF are:

- Passband (PB) which defines the range of frequencies transmitted (ideally without attenuation), its width is expressed by the bandwidth (BW).
- Center frequency (f_c) defines the frequency at the middle of the passband.
- Insertion Loss (IL) describes the losses between the input and the output at f_c. It is the value of the transmission coefficient (represented by the black curve in Fig. 12) at f_c expressed in dB; values close to 0 dB are sought.
- Return Loss (RL) describes the portion of the signal that is reflected back by the filter. It is the highest value of the reflection coefficient (represented by the gray curve in Fig. 12) in the passband and it is expressed in dB; typical values are lower than -10 dB.
- Fractional bandwidth (FBW) is defined as the ratio between the center frequency (f_c) and the width of the passband.

The reader can refer to Fig. 12 where the example of a filter response is reported along with the parameters defined above.

A comprehensive review on microwave filter technology has been reported in [56, 57], whilst filter design concepts and practical aspects can be found in [58, 59]. The main building block of a bandpass filter is termed resonator. The number of resonators employed determines the order of the filter (a filter with N resonators is termed an N^{th}-order filter). For instance, Fig. 13 shows the structure of a 6^{th}-order filter having six resonators coupled one to the other. A resonator is characterized

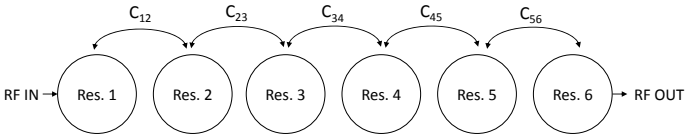


Fig. 13 Block diagram of a 6^{th}-order filter showing the building elements (resonators) and the energy transfer between them (couplings C_{ij})

by its resonant frequency which ultimately determines f_c of the filter. The unloaded quality factor (Q_u) of a resonator which describes how much energy is stored in the resonator compared to the dissipated energy determines the performance of the filter. For instance, in order for the filter to have a low IL, the resonators need to have a high Q_u. In the normal working condition of a filter, it is required that the energy is transferred from one resonator to the adjacent one (simplest case). The entity of this energy transfer, referred to as inter-resonator coupling, must be designed opportunely since it determines the bandwidth (i.e., also the FBW) of the filter. From the theory of coupled resonators, it is possible to demonstrate that the higher the coupling (also termed inter-resonator coupling) the wider the FBW [60]. The Q_u together with the FBW determines the IL of the filter, as shown by the formula [61]:

$$IL = 10\log(e)\frac{1}{FBW}\frac{1}{Q_u}\sum_{k=1}^{N} g_k, \tag{3}$$

where the coefficients g_k depend on the approximation used to synthesize the filter (Chebyschev, Butterworth, etc.,) and N is the order of the filter [62].

According to this formula, when the quality factor is fixed, wideband filters achieve lower IL compared to narrowband filters. This means that in WB and UWB filters, Q_u is not a critical aspect. Vice versa, narrowband filters exhibit high IL when the quality factor of the resonator is low.

Such considerations are independent of the technology employed to implement the filter. However, the technology used to implement a bandpass filter does have an influence on the quality factor of a resonator and the maximum bandwidth that can be obtained. Indeed, while it is usually not a problem to achieve low coupling coefficients between resonators (which results in narrowband filters) it is not always possible to obtain strong couplings between resonators which would lead to a wide and ultra-wide band filter response. This is the case for instance with dielectric-based resonators, often employed in order to shrink the size of the filter. In such resonators, where the electric field is strongly confined in the dielectric material the coupling of energy with adjacent resonators are difficult to obtain and so are strong couplings (wide and ultra-wide band filters cannot be obtained) [63]. This effect becomes more and more evident as the dielectric constant of the material increases. The considerations expressed so far are basic concepts valid for static filters but hold true also for tunable (or reconfigurable) filters that are key elements

for tunable transceivers and will be described in detail below. The discussion will start with narrowband filters and continue with the description of wideband and UWB bandpass filters, where the latter may have adaptive notches in their response. A notch is a narrowband bandstop filter. The latter typology of filter aims to block narrowband interferences at the front-end stage to increase reliability by preventing the saturation of the LNA (the reader may refer to Fig. 11 where the building blocks of RF receivers are shown).

4.2.1 Tunable Bandpass Filters

The tunability of a bandpass filter, for the most common cases, can be implemented in terms of center frequency and/or bandwidth. Their function enables dependability since the transmission of the signal is shifted to the most suitable frequency band at the time of the transmission. This is useful for example, in case of interferences from external sources which may cause the LNA to saturate. In other words, BPFs enable the implementation of frequency diversity schemes (see Sect. 3.1.1). The tuning elements employed for this purpose are usually switches and variable capacitors, to mention two representative examples. When switches are employed, the result is a filter response with a finite number of configurations, i.e., realizing *discrete reconfigurability*, given the nature of switches of having only two possible states.

When variable capacitors are employed, instead, the result is a continuous variation of the filter response, i.e., a *continuous reconfigurability* is achieved.

Both switches and variable capacitors can be realized in different technologies, e.g., they can be mechanical-, magnetic-, MEMS-, or semiconductor-based. The technology determines the switching speed which has a direct impact on the timeliness of the system. As an example, pin diodes provide very high switching speed and reliability (since the technology is very well established) therefore they are usually preferred. However, a trade-off between performance and power consumption must be observed. Indeed, the high current required by pin diodes (order of few mA) might have a negative effect on the availability of a wireless node due to excessive power consumption. MEMS represent a valid alternative, although researchers have been working in the last decade toward more reliable solutions for this technology [64].

The highest possible frequency/bandwidth shift is termed tuning range. Tunable filters that have a wide tuning range are very appealing since they can execute the function of several filters with a single device which reduces the size of the system. The main challenge in tunable filters is achieving a wide tuning range and preserving high performance (a high quality factor value within the entire tuning range, which ultimately results in maintaining IL and selectivity performance of the filter at an acceptable level). As already introduced, resonators are the building blocks of BPFs. By introducing a tuning element in a resonator that allows one to change its resonant frequency, it is possible to control the center frequency of the filter implemented with this resonator [65]. Similarly, if tuning elements are used to modify the coupling between resonators, the bandwidth of the filter is tuned. Such techniques are valid from NB up to UWB filters. However, for WB and UWB filters employed in IoT

systems, another type of tunability finds application, namely tunable notch filters which are presented in the next section.

4.2.2 Tunable Notch Filters

The last case of tunability presented in this section is that of WB and UWB filters with a reconfigurable notch in their response, referred to as tunable notch filters. UWB has emerged as a fast growing technology, however, a major impediment to the employment of UWB systems is the issue of narrowband interference that might exist in the same spectrum region. In UWB transceivers, this results in the saturation of the LNA. For these reasons, UWB filters able to reject specific frequencies are being developed [47]. Tunable notch filters contribute to achieve dependability as they enable notches (also more than one at the same time) in order to mitigate interference in a dynamic way. Depending on the way they are implemented, they can have a *discrete or a continuous reconfigurable* behavior which means they employ switches or variable capacitors, respectively [66, 67]. The attenuation of the interference provided by a notch usually ranges from -15 to -20 dB whilst IL is usually not an issue in such filters given their wide FBW (e.g., in [66] IL is better than -1.1 dB). The latter statement can be verified referring to Formula (3) where the FBW appears in the denominator. Thanks to the fact that IL is not an issue, such filters can be implemented in technologies that usually provide low Q_u, for instance based on silicon processes which is a good candidate for integrated UWB RF front-end modules [66]. This represents a big benefit since, technologies that provide high Q_u (e.g., waveguide or cavities in general) are not suitable for IoT applications due to the unreasonable dimension of their hardware (e.g., a standard waveguide for 2.4 GHz signals has a cross-section of 86×43 mm).

4.3 Antennas

As presented in Sect. 4.1, antennas are indispensable devices that enable IoT nodes to communicate with one another [68]. Antennas can be defined as a one port network passive device enabling the transition of electromagnetic waves from a guided wave to a free-space wave, or vice versa. Antennas have a strong impact on the dependability of a wireless communication link because of their direct influence on two main impairments, the multipath propagation and the interference from other radio sources. Antennas that are able to dynamically reconfigure/adapt their behavior by modifying one or more of their characteristics (frequency band, radiation properties, polarization, etc.,) enable therefore a higher reliability and timeliness [69, 70]. The directivity of an antenna can for instance, influence the fading distribution by suppressing strong multipath components (see "antenna diversity" in Sect. 5.3) or block interfering signals. Furthermore, increasing the antenna bandwidth contributes to increase the nodes' reliability by offering frequency diversity, see Section 3.1.1.

This section is devoted to an overview of antenna aspects and potentials in wireless IoT systems. Following the hardware design space shown in Fig. 9, challenges for increasing the antennas bandwidth are briefly highlighted in Sect. 4.3.1 while the antenna adaptability is discussed in Sect. 4.3.2. The main parameters used to characterize antennas are [71]:

- Antenna efficiency: is the ratio of the total power radiated by an antenna to the power delivered to the antenna.
- Directivity: is a measure of how directional an antenna is compared to an isotropic source (an isotropic antenna has zero directivity). The directivity characterizes the radiation properties of the antenna.
- Radiation pattern: is a mathematical function or graphical representation of the antenna radiation properties (gain, directivity, etc.,) as a function of space coordinates.
- Bandwidth: is defined as the difference of two frequencies on either side of the operating frequency (f_0) at which the antenna can radiate/receive energy.
- Polarization: is defined as the plane where the electric field oscillates while propagating. An antenna is called to be linearly (i.e., vertically or horizontally) polarized if its electric field is perpendicular or parallel to the Earth's surface. Circular or elliptical polarization occurs if an antenna electric field propagates in all planes (vertical, horizontal, and in between planes).

In order to guarantee an efficient transmission/reception, all aforementioned characteristics must be examined and suitably designed for an antenna to work properly. For example, the impedance between the antenna and the transmission line should be matched to obtain a maximum power transfer and thus radiation. It is usually said that an antenna is matched at a return loss of more than 10 dB over its operating frequency band, an example of an RL response can be seen in Fig. 12. Similarly, polarization matching is required for efficient transmission/reception as a vertically polarized antenna is not able to communicate with a horizontally polarized antenna. A detailed review of antenna design concepts and practical aspects can be found in [72–74].

4.3.1 Antenna Bandwidth: Towards Higher Bandwidth

Antennas can be classified in terms of bandwidth into narrowband, wideband, and ultra-wideband antennas (see Sect. 2.2). The theory of narrowband antennas have reached a certain level of maturity where many off-the-shelf products for IoT applications are already available. Wire and microstrip antennas are widely used in narrowband IoT nodes as they can be easily designed to operate at a predefined center frequency and bandwidth with low profile, cost, and ease of integration on printed circuit board materials. Narrowband antennas are typically resonant devices operating at a single resonance frequency at a time with a certain bandwidth, quality factor (see Sect. 4.2), and radiation efficiency. Chu and Harrington [75–77] investigated the fundamental limitation of electrically small antennas for achieving a broader

impedance bandwidth. It was found that to increase the antenna bandwidth, its quality factor has to be reduced, and thus the antenna radiation efficiency will degrade as well. In addition, the reduction in the antenna size will lead to a rapid increase in the antenna quality factor and thus limiting its bandwidth. It can be concluded that the antenna quality factor, bandwidth, efficiency and size are related and a trade-off between them is unavoidable to achieve an optimal design. Several techniques have been proposed to increase the antenna bandwidth. For instance, thickening the wire antennas and increasing the substrate thickness of microstrip antennas can lead to improvements of the impedance bandwidth by a few percents. Another way to obtain wideband antennas is by overlapping two or more resonant parts operating at their own resonances or by introducing multiple resonances within the same structure such as U-slot microstrip antennas, parasitic microstrip antennas, and stacked microstrip antennas [78]. UWB antennas can be designed by a combination of one or more of the following techniques [79]: electrically small antennas, frequency-independent antennas, multiple resonance antennas, traveling wave structures, and self-complementary antennas.

The main design challenges in UWB antennas for IoT applications are to realize a high radiation efficiency, linear phase, low dispersion, large bandwidth, compact size, and compatibility with integrated circuits. While the radiation of high power is allowed in narrowband applications, UWB transceivers' transmission power is below the noise floor level (in fact below $-41.3\,\mathrm{dBm/MHz}$ as shown in Sect. 5.2) which requires antennas with high radiation efficiency, i.e., conductor and dielectric losses have to be minimized, while maintaining good impedance bandwidth matching. In general, UWB transceivers use impulse radio signals for communication (see Sect. 3.1). The high reflection coefficient in ultra-wideband antennas can lead to a nonlinear phase and thus to a high distortion in the transmitted signal. In comparison to narrowband antennas where phase is considered constant, UWB antennas' radiation properties are frequency-dependent. The distortion might make it impossible to recover the transmitted pulse at the receiver. This reduces the IoT nodes dependability, as more computational processing power is required to recover the distorted pulse.

A detailed review of UWB antennas, potentials, and challenges can be found in [80–82]. An example of a UWB antenna with a continuously tunable and independent notch filters integrated on the same structure suitable for IoT applications can be found in [83]. In [84], a planar UWB antenna with an improved radiation performance versus frequency suitable for IoT applications is presented.

4.3.2 Antenna Adaptability/Reconfigurability

In general, a reconfigurable antenna can be realized by an antenna or an array of antennas and is capable of adapting its behavior by modifying one or more of its parameters (operating frequency, bandwidth, and radiation properties) in real time via electrical, mechanical, or other means. This section focuses on the reconfigurability

of the antennas' radiation properties (directionality), which contributes to spatial diversity utilization (see Sects. 3.1.1 and 5.3.6).

Reconfigurability of antenna radiation properties considers steering the main beam of directional antennas towards the desired direction of communications and, in the most advanced cases, nulls in the direction of unwanted signals. This is equivalent to spatial filtering to reduce interference and thus the IoT node reliability is improved. However, the use of reconfigurable antennas can degrade availability and timeliness when large numbers of switches, phase shifters and/or complex signal processing are used. The reduced battery lifetime might impair the availability of the sensor node drastically, as reconfigurable antennas' energy consumption depends on their complexity. Also, the IoT nodes timeliness might degrade drastically because of the delays caused by complex signal processing and/or mechanical antenna rotation. The main challenges in designing reconfigurable directional antennas lies in the system complexity (expertise is needed in different areas such as in antenna design, feeding networks, signal processing, complex measurement, and steering/beam-forming algorithms), cost, size, and power consumption.

As indicated in Fig. 15, several realizations of reconfigurable directional antennas (depending on their level of adaptability) are existing. The classical approach is *switchable antennas*, which are switching between antennas with fixed radiation patterns. To focus the power in another direction, the antenna has to be either mechanically rotated or a different antenna element has to be electronically switched on or off. Figure 14 shows an example of an electronically switchable directional antenna system for UWB-based IoT applications. It consists of four directional UWB antennas and an RF switching network controllable via two GPIO ports.

The disadvantage of switchable directional antennas is the limited degrees of freedom and typically also a higher form factor. An alternative is *smart antennas*, consisting of an array of several antenna elements and additional signal processing capabilities, which results in more degrees of freedom, like directional and adaptive beamforming. This advantage typically comes at the cost of high computational power for the adaptive excitation of the antenna elements (which is typically done in a microcontroller or digital signal processor), as well as a bigger size of the antenna system. In the simplest implementation of a smart antenna, the phase shifts of each antenna element to achieve a certain beam are preprogrammed in the memory of the processing unit (*switched smart antenna*). In a more adaptable system, the beam can be formed dynamically depending on the current interference sources and environment (*adaptive smart antenna*). In the most adaptable and cognitive cases, several adaptive smart antennas are necessary to exploit multipath components and achieve highest diversity gain (*MIMO*), as described in Sect. 3.1.1.

Additional challenges arise when ultra-wideband antennas are employed as radiating elements to form an array. As mentioned in the previous section, ultra-wideband antennas are frequency-dependent devices. Thus, the design of frequency-independent feeding networks is required to guarantee the IoT nodes' availability. The design of phase shifters and beam-forming algorithms becomes more challenging with increasing bandwidth. A detailed review on reconfigurable directional antennas can be found in [86–88].

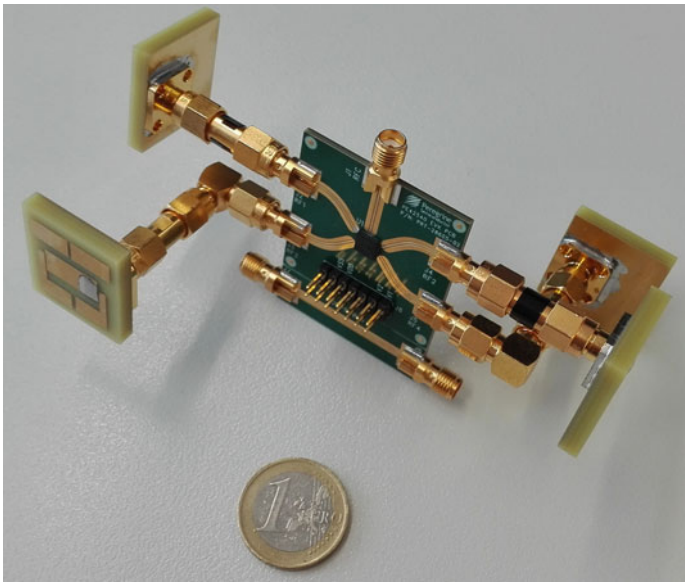


Fig. 14 Switchable antenna system for UWB-based IoT applications [85]

For more details about using switchable and smart antennas in wireless IoT networks and their positive impact on the dependability of IoT systems please refer to Sect. 5.3.6.

5 Networking

The previous sections described the wireless propagation channel and its characteristics focusing on a pair of communicating devices (i.e., considering only a single transmitter/receiver and the necessary hardware implementation). In most wireless systems and IoT applications, however, a large number of devices is forming a *network* in which nodes can broadcast information to several nodes, or in which pairs of nodes concurrently communicate with each other at the same time. This section, hence, investigates how multiple devices sharing the same medium can communicate in a dependable fashion. Section 5.1 first reviews the tasks of the medium access control layer in the context of IoT systems, and discusses its impact on key dependability attributes such as timeliness, availability, and reliability. Thereafter, Sect. 5.2 gives an overview of the wireless communication technologies used to build IoT applications to date, and highlights how the community has, so far, mostly focused on lower bandwidth. Finally, Sect. 5.3 introduces the design space in the networking domain, and shows that there is still plenty of room to improve the dependability of IoT systems by increasing the bandwidth and adaptability.

5.1 Impact of Medium Access Control on Dependability

Wireless communications are inherently broadcast. Therefore, the wireless channel has to be effectively shared between devices to ensure a dependable data transfer. Towards this goal, the medium access control (MAC) sub-layer of the data link layer[3] plays a key role, as it controls the access of the devices to the shared medium and is hence responsible for avoiding collisions and interference between multiple nodes. Furthermore, as IoT devices are typically resource-constrained and operating on battery, the MAC layer has often also the responsibility to minimize the time in which the radio transceiver of a device is listening for incoming packets in order to increase the availability of the system. The remainder of this subsection briefly reviews the role of the MAC layer with respect to the three dependability attributes introduced in Sect. 1: availability, timeliness, and reliability.

Availability. The majority of IoT applications employs battery-powered devices embedding a radio transceiver. The latter is a power-hungry component (typically, by far the most power-consuming component in a wireless sensor node) and it is important to minimize its active time to guarantee energy-efficient operations. Traditionally, this task is fulfilled by *duty-cycling* MAC protocols, that efficiently control the time in which a radio transceiver is turned on and off and increase the energy efficiency of the system and hence its availability. A large group of researchers has worked intensively on the design of energy-efficient duty-cycling MAC protocols since the early 2000s [5, 89, 90]. Influential examples are Sensor-MAC (S-MAC) [91], Timeout-MAC (T-MAC) [92], and X-MAC [93].

Timeliness. Periodically turning off the radio to save energy may increase availability, but may at the same time also have a strong impact on the ability of an IoT system to meet timeliness requirements. As an example, in vehicular communications, it is essential that information about the road conditions are received in a timely manner from a central server. Similarly, in a smart parking application, the end user may demand updates about the current parking situation from a smartphone in real time. This requires resource-constrained wireless sensors deployed in the streets to be reactive to such requests and hence to poll for incoming packets quite often (i.e., to frequently turn on the radio). This poses a sort of *catch-22* dilemma between low latency and high availability when meeting the application requirements—a very well-known problem that the research community is still trying to properly address [94]. For applications that clearly privilege only one of the two requirements (e.g., high availability over low latencies), quite a number of solutions have been proposed, such as Dozer [95] and BailighPulse [96].

Reliability. The MAC protocol also plays a crucial role in providing a high delivery rate and in satisfying minimum reliability requirements imposed by the application. In particular, the medium access control layer is often responsible for recovering

[3] Second layer in the OSI reference model.

from transmission errors and avoiding collisions caused by interference. Sources of interference can be internal to the network of interest or external [97]. External interference is traditionally caused by colocated wireless devices or appliances that radiate electromagnetic energy in the frequency bands used by the network of interest. Internal interference, instead, is caused by concurrent transmissions of other wireless devices operating in the same network. An overview of the solutions investigated by the community to mitigate external interference can be found in [97]. To solve the internal interference problem between devices operating in the same network, MAC protocols can either adopt a *contention-based* or a *schedule-based* approach.

Schedule-based MAC protocols assign the medium exclusively to a specific set of wireless devices. Frequency-division multiple access (FDMA) protocols [98, 99] subdivide the available bandwidth into smaller bands and assign each device to one of these frequency bands to communicate. Although this approach minimizes interference and maximizes the bandwidth available for communications, it may not scale to dense networks. Time-division multiple access (TDMA) protocols [100, 101] employ the same frequency band for all transmissions, but split the time domain into several time-slots. A time schedule indicates which device(s) may transmit frames during a certain time slot: the larger the frame size and the number of wireless devices, the higher are the delays before a node can get access to the medium. Protocols based on code- division multiple access (CDMA) use the same frequency and time-slot throughout the network, but employ simultaneous transmission by means of orthogonal codes. Typically, multiple access schemes can also be combined into hybrid approaches: the time-slotted channel hopping protocol (TSCH) is an exemplary protocol using a combination of FDMA and TDMA [102].

In *contention-based* approaches, instead, the medium is shared by multiple devices simultaneously. The simplest example of contention-based protocols is the so-called "pure" or *unslotted* ALOHA [103], which allows each device to transmit packets as soon as data is available. However, as the transmission time can be chosen arbitrarily, it is likely to generate collisions when several transmitters want to communicate simultaneously. In a *slotted ALOHA* system, instead, the transmission is just allowed in specific time-slots. Each transmitter can pick one of these slots, i.e., it synchronizes with the beginning of a time-slot, and a collision can only occur if multiple devices are sending during the same time-slot. This reduces the probability of collisions and doubles the maximum achievable throughput of unslotted ALOHA [6]. ALOHA, however, is particularly inefficient in crowded channels. To reduce the number of collisions, carrier-sense multiple access (CSMA)-based protocols have hence been introduced [104]. CSMA is based on clear channel assessment (CCA), which senses the wireless channel to check if there is an ongoing transmission. If this is the case, the transmitter backs off and postpones its transmission. If instead, the channel is not occupied, the packet can be transmitted immediately. A clear advantage of CSMA is that it does not require a close coordination among nodes (such as time synchronization), which is typically the case for TDMA-based protocols. One of the biggest drawbacks of CSMA protocols, however, is the "hidden node" problem [105] that occurs when a node A is visible from another node B, but not from other nodes communicating to B. A countermeasure to overcome

this problem is to carry out a handshake with special request-to-send/clear-to-send (RTS/CTS) messages, as introduced by MACA [106]. By sending an RTS message, the transmitter signals that it has a packet to be sent and includes the duration T of the planned transmission. If the receiver is ready to receive the message, it answers with a CTS message and all nodes receiving this message postpone their transmissions to avoid generating collisions.

5.2 Impact of Bandwidth on Dependability

Today, most of the deployed wireless sensor networks (WSNs) and IoT applications are based on the IEEE 802.15.4 standard [107], which defines the physical (PHY) and medium access control layers for low-cost, low-power, and low-rate wireless personal area networks (WPANs). The last revision of the standard defines 19 different PHYs, most of which use globally reserved industrial, scientific, and medical (ISM) radio bands. Among those bands, the most popular and used ones are:

- 868–868.6 MHz (Europe), one available channel
- 902–928 MHz (America), 10 available channels, channel spacing: 2 MHz
- 2400–2483.5 MHz (Worldwide), 16 available channels, channel spacing: 5MHz

IEEE 802.15.4 is not the only wireless technology employing these frequencies. In recent years, quite a number of standards, addressing IoT systems, have been specified, such as Bluetooth Low Energy (BLE),[4] LoRa, SIGFOX, Z-Wave, and Wi-Fi HaLow, just to name a few. Each of these technologies has different strengths and fields of application. BLE, for example, has the advantage of being ubiquitous nowadays, as most commercial tablets, smartphones, and laptops support it. LoRa, SIGFOX, and Wi-Fi HaLow offer long-range communication over several kilometers. However, all these technologies share a common limitation: they are inherently narrowband communication technologies. This causes these systems to be highly susceptible to multipath fading (see Sect. 2.1) and cross-technology interference, which reduces throughput and leads to an increased amount of network traffic due to retransmissions [97, 108, 109].

A promising alternative to tackle these limitations is the shift towards higher bandwidth. The high bandwidth (allowing short pulses) results in beneficial properties such as a high immunity to multipath fading, a very good time-domain resolution allowing for precise localization and tracking, as well as possible high data rates. The IEEE 802.15.4 working group recognized this potential and published in 2007 the IEEE 802.15.4a amendment. The latter specifies additional PHYs to add a ranging capability with an accuracy of one meter or higher, an extended communication range, as well as improved robustness and mobility support as compared to the IEEE 802.15.4-2003 standard [110]. One of the added PHY layers

[4] BLE is marketed as Bluetooth Smart and was originally introduced as Wibree by Nokia. It is merged by the Bluetooth special interest group (SIG) into the Bluetooth Core Specification v4.0.

is the impulse radio ultra-wideband (IR-UWB) technology. As briefly introduced in Sect. 2.2, UWB-based devices spread the signal power over a wide bandwidth ($\geq$ 500 MHz) yielding extremely low-power spectral density and, as a consequence, reduce interference to other systems. In February 2002, the Federal Communications Commission (FCC) allocated the 3.1–10.6 GHz frequency band for unlicensed use with a maximum equivalent isotropically radiated power (EIRP) of -41.3 dBm/MHz, which is also the limit for unintentional radiators (e.g., TVs and monitors). Gradually, also other countries—with slight differences to the FCC spectrum mask—defined their own UWB regulations [111].

In the wireless sensor networks and IoT research communities, UWB communication systems have drawn significant interest in the past, but without making the breakthrough and without finding the way into off-the-shelf consumer products [112]. For this reason, as shown in the next subsection, most work on dependable networking has been focusing on narrowband technologies only. However, this may change in the coming years, especially after the publication of IEEE 802.15.4a standard and the commercialization of the first low-cost IEEE 802.15.4-compliant UWB transceiver, the DecaWave DW1000 [113]. These two key factors, together with the outstanding localization performance of UWB technology [114, 115] (which was also proven in harsh environments such as mines [116]) aroused the enthusiasm for UWB technology and its potential for dependable IoT applications.

5.3 Impact of Networking Design Space on Dependability

The design space of the networking section holds, as before, *adaptability* on the x-axis and *bandwidth* on the y-axis as shown in Fig. 15. The x-axis is subdivided into *static*, *switchable*, *adaptive/reactive*, *cognitive/predictive*, whereas the y-axis is subdivided into *narrowband* and *ultra-wideband*. The presented design space covers state-of-the-art technologies in the networking domain while presenting their impact on the dependability of a wireless system. In particular, each of the techniques shown in Fig. 15 is treated separately and analyzed in relation to bandwidth and adaptability, as well as exemplary MAC protocols are discussed. Finally, it is highlighted that there is significant room for future work in the region of higher bandwidth and higher adaptability (indicated by the red rectangle) and it is argued that future research should address this area.

5.3.1 Static Channel Assignment

When designing a wireless system, it has to be defined which frequency bands and wireless channels should be used for communication. In the simplest case, a static channel is assigned during the deployment phase and is not to be changed throughout the lifetime of a network. The selection of the channel may take communication range or surrounding interference into account. As indicated in Fig. 15, static

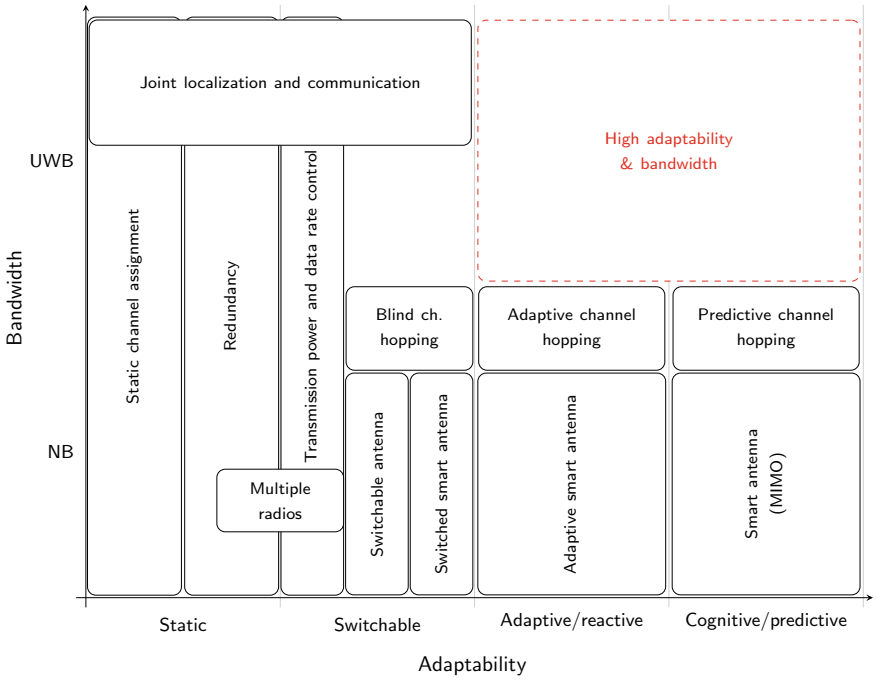


Fig. 15 Networking design space. State-of-the-art techniques used in the networking domain

channel assignment in a wireless system is used by all technologies regardless of their bandwidth. Still, the external interference in the narrowband case and the usage of preamble codes in ultra-wideband motivate a few important remarks.

Narrowband. Especially in the crowded 2.4 GHz ISM band, several technologies and devices are coexisting in the same frequency range [108]. In the likely case of concurrent wireless transmissions in the surroundings (e.g., caused by a Wi-Fi access point or a Bluetooth-enabled smartphone), the quality of the statically defined channel may decrease significantly and, as a result, the link could become highly unreliable [97]. The only degree of freedom to increase the reliability within this approach is the selection of a channel with minimal overlap to other technologies. For example, several WSN applications employ IEEE 802.15.4 channel 26 to escape at least the interference of surrounding Wi-Fi devices operating in the 2.4 GHz ISM band [117, 118].

Ultra-wideband. UWB-based applications typically use a static channel. Nevertheless, as specified by the IEEE 802.15.4 UWB standard, several networks can coexist on the same channel using different preamble codes. The preamble code is sent before the physical header and data field of the IEEE 802.15.4 packets for synchronization and channel estimation. The low cross-correlation between several preamble codes

allows simultaneously operating networks. In the case of a decentralized multiple access technique, a smart way to assign preamble codes to the devices and networks is required. The following three code assignment approaches can be found in the literature [119, 120]:

1. Common code: the simplest principle is that just one code is used for all transmissions. If several nodes are transmitting simultaneously, collisions may occur.
2. Receiver-based code: each user has a unique receiving code. Therefore, the receiver has to monitor only its receiving code. Collisions are possible when several users try to transmit data to the same receiver.
3. Transmitter-based code: each user is assigned to a unique sending code. Collisions between different transmitters do not appear anymore, but it is necessary that the receiver knows with which code to listen.

Practical implementations often use a hybrid approach, such as common-transmitter (C-T) or receiver-transmitter (R-T) codes. These are then combined with the RTS/CTS mechanism to form the so-called MACA/C-T or MACA/R-T protocols [121]. In MACA/C-T, RTS/CTS messages are transmitted via the common code approach and the data via the unique sending code (transmitter-based). In MACA/R-T, a unique receiving and sending code is assigned to each node. RTS is sent with the destination's receiving code, whereas CTS and data are transmitted with the appropriate sending code.

5.3.2 Redundancy

Regardless of the employed bandwidth, a classical approach used to increase the dependability of a communication link is to mitigate the impact of errors in the data transmission by means of redundancy. The simplest example is to repeat the whole information multiple times (see time diversity concept in Sect. 3.1.1): this is for example, done by default in the IEEE 802.15.1 standard (Bluetooth), and also proposed in IEEE 802.15.4 (in the context of packet headers) to mitigate the impact of surrounding external Wi-Fi access points [122] . A more efficient approach is *forward error correction* (FEC), in which additional information is added to the original packet and is then used to detect or even correct possible errors. If the latter are corrected directly at the receiver, no packet retransmissions are necessary. However, embedding redundant information in a packet results in a larger overhead in terms of a longer transmission time, as well as additional time for encoding and decoding of the error-correcting code. Therefore, FEC is used when retransmission is costly or even impossible, i.e., in unidirectional or multicast communications.

Another way to introduce redundancy is the so-called *backward error correction* (BEC). In BEC, the additional message types acknowledgement (ACK) or negative acknowledgement (NACK) are used to inform the transmitter whether a packet is successfully received or lost, respectively. Depending on the implementation, the transmitter has to decide whether a packet should be retransmitted.

Basic FEC techniques are clearly static approaches, which is indicated in Fig. 15. Whereas, BEC techniques include some adaptive elements, since the amount of sent ACKs is depending on the quality of the link and the reception rate. Furthermore, several BEC algorithms are, for example, adaptively determining the timeout before re-transmitting a packet [123]. In the bandwidth domain of the design space, the redundancy block covers the whole axis, which indicates that the usage of redundancy techniques is in theory not influenced by the bandwidth.

5.3.3 Multiple Radios

As indicated in Sect. 4.1, a cost- and space-intensive solution is to increase the number of transceivers embedded on a wireless device to improve the dependability of a network. In this regard, it has to be differentiated between having transceivers using the same radio technology (as an example, Draves et al. [124] use two 802.11 transceivers per node, where the individual radios are tuned to different, non-interfering channels) and having independent radio technologies (as an example, the BTnode from ETH Zurich [125] is equipped with a Bluetooth radio and a low-power sub-GHz ISM band radio that can be operated simultaneously or independently powered off).

Although the introduction of Bluetooth Low Energy may have diminished the need for a low-power technology in addition to Bluetooth, the concept of dual- or multi-radio is widely used in the research community. As also indicated in the design space in Fig. 15, so far, this technique was mainly used in narrowband applications. However, because of its outstanding ranging performance, UWB radios are increasingly used as a localization technology, mainly in combination with an additional narrowband communication module (see Sect. 5.3.7). Individually switching between different radio models requires a certain level of adaptability. For this reason, this technique is placed in the cross-section between static and switchable in the design space.

5.3.4 Transmission Power and Rate Control

The control of transmission power and data rate is a useful tool to manage the interference level and network reliability. Transmission power control is a well-known technique in narrowband applications. Lin et al. [126] have presented ATPC, a lightweight algorithm to adapt the transmission power and minimize internal interference in a wireless sensor network. Similarly, Shen et al. [119] and Cuomo et al. [127] have also shown the importance of transmission power control for higher bandwidth. The technique, therefore, covers the whole y-axis of the design space depicted in Fig. 15.

It is important to highlight that transmission power control has also an impact on network availability, as it often allows to decrease the transmit power levels of the devices in the network to the minimum amount necessary to reach the intended

receiver(s). Additionally, in IEEE 802.15.4-compatible UWB transceivers, physical layer parameters such as pulse repetition frequency (PRF) or preamble length can be changed, depending on the application and environment [128].

5.3.5 Channel Hopping

Instead of transmitting on the same highly congested frequency band, a device can hop across different channels. Channel hopping reduces fading by means of frequency diversity [129]. This assumes that the channels are spaced apart by more than the coherence bandwidth to reduce the probability of a simultaneous deep fade at both channels (see Sect. 2.1 for further details). Depending on their level of adaptability, one can differentiate between blind, adaptive, and predictive hopping [97].

Channel hopping is essentially exploiting frequency diversity, and implies that a higher bandwidth w.r.t. the static channel approach is used, which is also mapped in Fig. 15 accordingly. It is important to highlight that accurate time synchronization in the network is traditionally necessary in order to allow all nodes to hop in unison.

Blind channel hopping. In blind hopping, the wireless nodes follow a pseudo-random sequence to hop continuously between the available channels, and there is no prior knowledge of the link quality of the channels necessary. If not all channels are highly congested, the average interference level should be significantly decreased. However, if most of the channels are crowded, blind channel hopping is ineffective. Since the IEEE 802.15.4e amendment was released in 2012 [102], also the IEEE 802.15.4 standard explicitly supports channel hopping. The employed protocol is called TSCH, and supports time-slotted access together with channel hopping. Other standards that make use of continuous hopping are WirelessHART [130] and ANT+ [131].

Adaptive channel hopping. To avoid hopping back to congested channels, a device can store and remember the number of the congested channel and hop accordingly in the next hopping cycle. This technique is also called blacklisting. An adaptive version of TSCH was presented by Du et al. [132]. Since version 1.2 (2003), also Bluetooth supports adaptive hopping by avoiding the use of crowded frequencies, called adaptive frequency hopping (AFH). The challenge is to identify whether a channel is sufficiently "good" or not. This requires an efficient and well-performing CCA, which is difficult to achieve for impulse radio ultra-wideband (see Sect. 5.3.7). A drawback of adaptive channel hopping is the necessity for sharing the list of blacklisted channels throughout the network.

Predictive channel hopping. Before blacklisting channels and adapting the hopping sequence to the interference in the surroundings, the wireless device has to first estimate the quality of available channels. This may require to periodically send or (attempt to) receive one or more sample packets for each frequency band. However, sampling interfered channels may cause significant packet loss or trigger a number of retransmissions that may significantly degrade performance. For this reasons, the

most desirable concept is to predict the deterioration of channel conditions beforehand. This is typically referred to as *predictive* or *proactive channel hopping* [97]. A fundamental role in this regard is played by channel quality estimation metrics that can detect an early degradation of the channel [133, 134], as well as by an efficient link quality ranking algorithm [135] and interference classification schemes [136, 137]. Although a number of predictive protocols have been proposed for wireless sensor networks operating in the 2.4 GHz band [138, 139], the main disadvantage of these approaches is that they heavily rely on high-rate and accurate energy detection, which is very costly in terms of energy consumption.

5.3.6 Antenna Diversity in the Networking Domain

The key idea behind antenna diversity is that the received signals of different antennas are uncorrelated, which is either achieved by spacing the antenna elements sufficiently far away from each other (see Sect. 3.1.1), or by making use of reconfigurable directional antennas (see Sect. 4.3.2).

The majority of wireless networks in use nowadays are still using omnidirectional antennas, i.e., the radio signal is transmitted in each direction equally (in the idealized case of isotropic radiation) and no other network user is allowed to transmit to avoid collisions, which results in a low spatial reuse and a reduced network capacity. For this reason, several research groups are investigating directional antennas. Advantages such as reduced contention and increased throughput [140], reduced interference [141], minimal packet error rate, as well as improved energy efficiency [142] were already shown for narrowband applications.

All the aforementioned work is using the classical approach of switchable antennas, which is the least adaptable version of directional antennas, as shown in Fig. 15. Antenna systems with a higher level of adaptability, such as smart antennas, and their design challenges are presented in Sect. 4.3.2. But the usage of switchable and smart antennas, despite their enormous potential, not just complicates the design of the physical layer in terms of size and computational complexity, but even more the design of the upper layers, e.g., the MAC layer [143]. Most MAC protocols are indeed using CCA to detect if a channel is free and packets can be transmitted. Also, the receiver has to sense the channel and, in the case of an incoming message, has to wake up the CPU. However, the receiver does typically not know the direction of an incoming message, which is why it either activates all antennas (and hence loses the advantage of directionality), or it activates the beams one after each other (which comes at the price of a higher power consumption and latency). This implies that the antenna beam has to be focused in the appropriate direction before the transmission and reception of packets in order to reach the highest quality of the communication link and escape interference. This task is even more challenging in dynamic and highly mobile networks with frequent node movements [144].

While setting up a network, users have no knowledge of when and in which direction they have to point their beam. Therefore, when using directional antennas a proper localization of each node is crucial. In this regard, several concepts are

presented in Sect. 3.2 and joint communication and localization is covered in more detail in the next subsection. Besides self-localization, wireless devices also need information about the position of the neighboring nodes in the network (neighbor discovery). Using directional antennas at the receiver and transmitter increases the complexity of this operation [143].

5.3.7 Ultra-Wideband MAC Protocols and Joint Localization and Communication

Several survey papers on existing MAC protocols for narrowband communication technologies have been published in the last decade [89, 90]. However, only limited work has been published about the influence of higher bandwidth on the design of low-power and reliable MAC protocols. As shown by Radunovic et al. [145], simply reusing MAC protocols that have been originally designed for narrowband systems might not be a good idea. Nevertheless, UWB MAC protocols can benefit from existing narrowband solutions, although the unique characteristics of ultra-wideband have to be properly addressed.

Because of their success in narrowband MAC protocols, it is a logical step to consider CSMA-based protocols also for ultra-wideband technologies. For this purpose, however, achieving accurate clear channel assessment is necessary (see Sect. 5.1). Typical narrowband receivers perform this by means of energy detection of carrier waveforms, so the channel is considered as clear as soon as the received signal strength is below a predefined threshold. Realizing this impulse radio UWB transceiver system is a challenging task, as the extremely low power density of a UWB spectrum causes an energy level that is typically below the noise floor. Hence, a conventional energy detection method based on a threshold does not work for UWB systems [146]. Consequently, a CSMA-based protocol without the ability to sense the channel results in a simple ALOHA-based protocol. As an alternative CCA procedure for UWB systems, the received preambles that are sent in IEEE 802.15.4 packets for synchronization and channel estimation can be used as an indicator for an ongoing transmission [111]. Such a preamble-detection-based CCA technique is presented by Qi et al. in [146]. Preamble symbols are inserted also in the header and payload parts of the IEEE 802.15.4 packet. This technique was adopted by the IEEE 802.15.4a standard.

Joint localization and communication. In many applications for the IoT, knowing the exact position of wireless nodes and their neighbors is a key aspect. Several measurements and sensor data collected from deployed wireless devices, indeed, just make sense if they include temporal and spatial information. However, RF-based narrowband localization technology cannot provide enough accuracy for most IoT applications. That is why current implementations use a radio technology for communication (e.g., BLE, Wi-Fi, and IEEE 802.15.4) and a different, more precise, technology for localization (e.g., ultrasound or light), which unfavorably affects the

form factor and costs of the devices. For example, Lazik et al. [147] use a combination of BLE and ultrasound to achieve decimeter-accurate localization.

Ultra-wideband technology can possibly provide a solution for both communication and localization purposes [148] to enable location-aware networking with a single wireless technology. Because of its physical properties, UWB technology has the ability to provide centimeter accuracy for ranging between nodes, and therefore, outclasses all of its narrowband competitors (see Sect. 3.2.2). The compatibility to the IEEE 802.15.4 standard makes it suitable for communication purposes. Still, UWB transceiver manufacturers typically motivate developers to use their chip either for communication or ranging, one at a time. But Alcock et al. [149] have already shown that also synchronous communication and positioning is possible using specialized ranging packets. They have modified the contention-based low-power MAC protocol *FrameComm* to make also use of distance measurements. This has the advantage that ranging does not consume additional energy and does not degrade the throughput, because existing messages are used for calculating the distance. Additionally to the communicating nodes, other devices in the network overhear the communication and can react with a ranging acknowledgement. In this way, ranging information to more than one node can be collected. Another location-aware UWB MAC protocol is presented in [150]: the proposed protocol (PMAC) is TDMA-based, distributed, and supports a dynamic network topology.

6 Conclusions and Future Work

The design space presented in the previous sections has shown the gradual shift of the research community towards highly configurable solutions targeting the vision of fully cognitive systems. This shift allows to design IoT systems that can satisfy the stringent requirements imposed by safety-critical applications on a large scale. Furthermore, the community also started to increase the bandwidth and exploit the resulting high time resolution for improved ranging and localization applications. Although the employment of ultra-wideband and highly configurable systems is challenging in terms of hardware requirements (as shown in Sect. 4), the research in the networking domain should expand to this area.

Indeed, despite the enormous amount of techniques and technologies that have been proposed so far, still significant work remains to develop energy-efficient fully cognitive radios and protocols. Especially in the networking domain, as shown in Fig. 15, there is still significant room for future work in the region of higher bandwidth and higher configurability, which would push the dependability of IoT communication even further. Towards this goal, the author's research is aiming at employing location-resolved models of the environment together with adaptive ultra-wideband radio front-ends (i.e., tunable filters and antennas) to support low-power operation and to increase reliability on a large scale [2]. By mapping the problem to a model-predictive control system that includes the adaptable radio front-ends, physical-layer signal processing for environment modeling and localization, and communication

protocols for distributed control of the radio transceivers, one can gain control over and satisfy the required dependability attributes. The low-cost and low-power UWB transceiver DecaWave DW1000 [113] enables the use of UWB technology in IoT deployments. In [151] this device was analyzed in terms of localization performance in comparison to a high fidelity measurement system. As an initial step, the authors proposed the application of a switchable UWB antenna system to enhance IoT communication and localization [85] and showed its potential for multipath-resolved positioning [41]. Future work will focus on extending the capabilities of this RF front-end using antenna arrays to decrease the degrees of freedom and provide a higher form factor. Furthermore, the antennas will be combined with tunable filters.

As shown in Sect. 4, in the hardware domain, the efforts are leading towards SDR/CR-based transceivers for IoT nodes [67, 152]. These nodes will then operate by dynamically sensing the frequency spectrum, finding the available bands in a target spectral range, and then transmitting immediately without introducing harmful interference to other nodes. Consequently, this allows efficient energy use of RF devices, which results in a high availability by maximizing the lifetime of IoT nodes [46].

This book chapter illustrated the complexity of providing dependable communication and localization for future IoT applications and emphasized the need for close cooperation between different domains, like signal processing, microwave engineering, and networking. Furthermore, aiming towards highly configurable and cognitive techniques, as well as higher bandwidth, was motivated by highlighting that current and past research shows a substantial gap in that area.

Acknowledgements This work was performed within the LEAD-Project "Dependable Internet of Things in Adverse Environments" funded by Graz University of Technology, Austria, and partly within the "Kalium Home Monitoring" project funded by the Austrian Research Promotion Agency (FFG), Austria.

References

1. Avižienis, A., Laprie, J.-C., Randell, B., Landwehr, C.: Basic concepts and taxonomy of dependable and secure computing. IEEE Trans. Dependable Secur Comput. **1**(1), 11–33 (2004)
2. Boano, C.A., Römer, K., Bloem, R., Witrisal, K., Baunach, M., Horn, M.: Dependability for the internet of things—from dependable networking in harsh environments to a holistic view on dependability. e&i Elektrotechnik und Informationstechnik **133**(7) (2016)
3. You, C.-W., Wei, C.-C., Chen, Y.-L., Chu, H.-H., Chen, M.-S.: Using mobile phones to monitor shopping time at physical stores. IEEE Pervasive Comput. **10**(2), 37–43 (2011)
4. Martella, C., Miraglia, A., Cattani, M., van Steen, M.: Leveraging proximity sensing to mine the behavior of museum visitors. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications (PerCom) (2016)
5. Langendoen, K.: Medium access control in wireless sensor networks, vol. 2, pp. 535–560 (2008)
6. Molisch, A., Wireless Communications. Wiley (2011)
7. Rappaport, T.S., Wireless Communications: Principles and Practice. Prentice Hall (2001)

8. Parsons, J.D.: The Mobile Radio Propagation Channel. Wiley (2000)
9. Gallager, R.G.: Stochastic Processes: Theory for Applications. Cambridge University Press (2013)
10. Molisch, A.F.: Ultra-wide-band propagation channels. Proc. IEEE **97**(2), 353–371 (2009)
11. Ultrawideband propagation channels-theory, measurement, and modeling. IEEE Trans. Veh. Technol. **54**(5), 1528–1545 (2005)
12. Proakis, J.G., Salehi, M., Zhou, N., Li, X.: Communication Systems Engineering, vol. 94. Prentice Hall, New Jersey (1994)
13. Barry, J.R., Lee, E.A., Messerschmitt, D.G.: Digital Communication. Springer Science & Business Media (2004)
14. Simon, M.K., Alouini, M.-S.: Digital Communication Over Fading Channels, vol. 95. Wiley (2005)
15. Alamouti, S.M.: A simple transmit diversity technique for wireless communications. IEEE J. Sel. Areas Commun. **16**(8), 1451–1458 (1998)
16. Cheun, K.: Performance of direct-sequence spread-spectrum rake receivers with random spreading sequences. IEEE Trans. Commun. **45**(9), 1130–1143 (1997)
17. Nee, R.V., Prasad, R.: OFDM for Wireless Multimedia Communications. Artech House, Inc. (2000)
18. Gast, M.: 802.11 Wireless Networks: The Definitive Guide. O'Reilly Media, Inc. (2005)
19. Dahlman, E., Parkvall, S., Skold, J.: 4G: LTE/LTE-Advanced for Mobile Broadband. Academic Press (2013)
20. Bohge, M., et al.: Dynamic resource allocation in ofdm systems: an overview of cross-layer optimization principles and techniques. IEEE Netw. **21**(1), 53–59 (2007)
21. Hanzo, L.L., et al.: MIMO-OFDM for LTE, WiFi and WiMAX: Coherent Versus Non-coherent and Cooperative Turbo Transceivers, vol. 9. Wiley (2010)
22. Malik, W.Q., Allen, B., Edwards, D.J.: Bandwidth-dependent modelling of smallscale fade depth in wireless channels. IET Microwaves Antennas Propag. **2**(6), 519–528 (2008)
23. Romme, J., Kull, B.: On the relation between bandwidth and robustness of indoor UWB communication. In: 2003 IEEE Conference on Ultra Wideband Systems and Technologies (2003)
24. Scholtz, R.: Multiple access with time-hopping impulse modulation. In: 1993 Military Communications Conference, MILCOM '93, Conference Record Communication on the Move. IEEE (1993)
25. Win, M.Z., Scholtz, R.A.: Ultra-wide bandwidth time-hopping spread-spectrum impulse radio for wireless multiple-access communications. IEEE Trans. Commun. **48**(4), 679–689 (2000)
26. Rushforth, C.: Transmitted-reference techniques for random or unknown channels. IEEE Trans. Inf. Theory **10**(1), 39–42 (1964)
27. Witrisal, K., Leus, G., Janssen, G.J., Pausini, M., Trösch, F., Zasowski, T., Romme, J.: Noncoherent ultra-wideband systems,. IEEE Signal Proc. Mag. **26**(4) (2009)
28. Chen, Y., Wang, B., Han, Y., Lai, H.Q., Safar, Z., Liu, K.J.R.: Why time reversal for future 5G wireless? [perspectives]. IEEE Signal Proc. Mag. **33**(2), 17–26 (2016)
29. Mitola, J.: The software radio architecture. IEEE Commun. Mag. **33**(5), 26–38 (1995)
30. Haykin, S.: Cognitive radio: brain-empowered wireless communications. IEEE J. Sel. Areas Commun. **23**(2), 201–220 (2005)
31. Haykin, S.: Cognitive Dynamic Systems: Perception-action Cycle, Radar and Radio. Cambridge University Press (2012)
32. Mok, E., Retscher, G.: Location determination using WiFi fingerprinting versus WiFi trilateration. J. Locat. Based Serv. **1**(2), 145–159 (2007)
33. Meissner, P.: Multipath-Assisted Indoor Positioning. Ph.D. Dissertation, Graz University of Technology (2014)
34. Durrant-Whyte, H., Bailey, T.: Simultaneous localization and mapping: part i. IEEE Robot. Autom. Mag. **13**(2), 99–110 (2006)
35. Thrun, S., Leonard, J.J.: Simultaneous localization and mapping. In: Springer Handbook of Robotics. Springer, pp. 871–889 (2008)

36. Leitinger, E., Meissner, P., Lafer, M., Witrisal, K.: Simultaneous localization and mapping using multipath channel information. In: Proceedings of the IEEE ICC-15 (2015)
37. Leitinger, E.: Cognitive indoor positioning and tracking using multipath channel information. Ph.D. Dissertation, Graz University of Technology (2016)
38. Kay, S.M.: Fundamentals of Statistical Signal Processing, Estimation Theory, vol. I, Prentice Hall (1993)
39. Witrisal, K., Leitinger, E., Hinteregger, S., Meissner, P.: Bandwidth scaling and diversity gain for ranging and positioning in dense multipath channels. IEEE Wirel. Commun. Lett. **PP**(99), 1–1 (2016)
40. Witrisal, K., et al.: High-accuracy localization for assisted living: 5G systems will turn multipath channels from foe to friend. IEEE Signal Proc. Mag. **33**(2), 59–70 (2016)
41. Rath, M., et al.: Multipath-assisted indoor positioning enabled by directional UWB sector antennas. In: IEEE International Workshop on Signal Processing Advances in Wireless Communications, SPAWC (2017)
42. Cruz, P., Carvalho, N.B., Remley, K.A.: Designing and testing software-defined radios. IEEE Microw. Mag. **11**(4), 83–94 (2010)
43. Pozar, D.: Microwave and RF Wireless Systems. Wiley, New York (2001)
44. Balasubramanian, S., Boumaiza, S., Sarbishaei, H., Quach, T., Orlando, P., et al.: Ultimate transmission. IEEE Microwav. Mag. **13**(1), 64–82 (2012)
45. Prata, A., Oliveira, A.S.R., Carvalho, N.B.: An agile digital radio system for UHF white spaces. IEEE Microwav. Mag. **15**(1), 92–97 (2014)
46. Kitsunezuka, M., Kunihiro, K., Fukaishi, M.: Efficient use of the spectrum. IEEE Microw. Mag. **13**(1), 55–63 (2012)
47. Gómez-García, R., et al.: Filling the spectral holes: novel/future wireless communications and radar receiver architectures. IEEE Microw. Mag. **15**(2), 45–56 (2014)
48. Kloc, M., et al.: Let's make them cognitive: cognitive radio technology applied to professional wireless microphone systems. IEEE Microw. Mag. **17**, 70–78 (2016)
49. Baylis, C., Fellows, M., Cohen, L., II, R.J.M.: Solving the spectrum crisis: intelligent, reconfigurable microwave transmitter amplifiers for cognitive radar. IEEE Microw. Mag. **15**(5), 94–107 (2014)
50. Maurer, L., Stuhlberger, R., Wicpalek, C., Haberpeuntner, G., Hueber, G.: Be flexible. IEEE Microw. Mag. **9**(2), 83–95 (2008)
51. Rofougaran, A.R., Rofougaran, M., Behzad, A.: Radios for next-generation wireless networks. IEEE Microw. Mag. **6**(1), 38–43 (2005)
52. Chastellain, F., Botteron, C., Farine, P.A.: Looking inside modern receivers. IEEE Microw. Mag. **12**(2), 87–98 (2011)
53. Wang, H., Dang, V., Liu, Q., Ren, L., Ren, L., et al.: An elegant solution: an alternative ultra-wideband transceiver based on stepped-frequency continuous-wave operation and compressive sensing. IEEE Microw. Mag. **17**(7), 53–63 (2016)
54. Nader, C.: et al. Wideband radio frequency measurements: from instrumentation to sampling theory. IEEE Microw. Mag. **14**(2), 85–98 (2013)
55. Mitola, J., Maguire, G.Q.: Cognitive radio: making software radios more personal. IEEE Personal Commun. **6**(4), 13–18 (1999)
56. Levy, R., Cohn, S.: A history of microwave filter research, design and development. IEEE Trans. Microw. Theory Tech. **32**(9), 1055–1067 (1984)
57. Hunter, I.C., Billonet, L., Jarry, B., Guillon, P.: Microwave filters—applications and technology. IEEE Trans. Microw. Theory Tech. **50**(3), 794–805 (2002)
58. Levy, R., Snyder, R., Matthaei, G.: Design of microwave filters. IEEE Trans. Microw. Theory Tech. **50**(3), 783–793 (2002)
59. Snyder, R.: Practical aspects of microwave filter development. IEEE Microw. Mag. **8**(2), 42–54 (2007)
60. Hong, J.-S.: Microstrip Filters for RF/Microwave Applications. Wiley (2011)
61. Bianchi, G., Sorrentino, R.: Electronic Filter Design and Simulation. McGraw-Hill (2007)

62. Matthaei, G., Jones, E., Young, L.: Microwave Filters, Impedance-Matching Networks, and Coupling Structures. Artech Microwave Library, North Bergen, NJ (1964)
63. Cameron, R., Kudsia, C., Mansour, R.R.: Microwave filters for communication systems. Wiley-Interscience, Hoboken, N.J. (2007)
64. Tazzoli, A., Peretti, V., Gaddi, R.: et al. Reliability issues in rf-mems switches submitted to cycling and esd test. In: 2006 IEEE International Reliability Physics Symposium Proceedings (2006)
65. Wong, P.W., Hunter, I.C.: Electronically reconfigurable microwave bandpass filter. IEEE Trans. Microw. Theory Tech. **57**(12), 3070–3079 (2009)
66. Wu, Z., Shim, Y., Rais-Zadeh, M.: Miniaturized UWB filters integrated with tunable notch filters using a silicon-based integrated passive device technology. IEEE Trans. Microw. Theory Tech. **60**(3), 518–527 (2012)
67. Pelliccia, L.: et al.: Compact ultra-wideband planar filter with RF-MEMS-based tunable notched band. In: 2012 Asia Pacific Microwave Conference Proceedings (2012)
68. Kraus, J.D., Marhefka, R.J.: Antennas for all Applications, 1st edn. McGraw-Hill (2002)
69. Chen, Z.N.: UWB antennas: design and application. In: 6th International Conference on Information, Communications & Signal Processing, vol. 2007, pp. 1–5. IEEE (2007)
70. Haider, N., Caratelli, D., Yarovoy, A.G.: Recent developments in reconfigurable and multiband antenna technology. Int. J. Antennas Propog. 1–14 (2013)
71. Balanis, C.A.: Antenna Theory: Analysis and Design (2016)
72. Gross, F.B.: Frontiers in Antennas, 1st edn. McGraw-Hill (2011)
73. Morishita, H., et al.: Design concept of antennas for small mobile terminals and the future perspective. IEEE Antennas Propag. Mag. **44**(5), 30–43 (2002)
74. Yang, T., Davis, W.A., Stutzman, W.L.: Fundamental-limit perspectives on ultrawideband antennas. Radio Sci. **44**(01), 1–8 (2009)
75. Harrington, R.F.: Effect of antenna size on gain, bandwidth, and efficiency. J. Res. Nat. Bur. Stand **64**(1), 1–12 (1960)
76. Chu, L.J.: Physical limitations of omni-directional antennas. J. Appl. Phys. **19**(12), 1163–1175 (1948)
77. McLean, J.S.: A re-examination of the fundamental limits on the radiation Q of electrically small antennas. IEEE Trans. Antennas Propag. **44**(5), 672 (1996)
78. Bhatnagar, M.: Broadband design of microstrip antennas: recent trends and developments. In: International Conference on Recent Advances in Microwave Theory and Applications (2008)
79. Adamiuk, G., Zwick, T., Wiesbeck, W.: UWB antennas for communication systems. Proc. IEEE **100**(7), 2308–2321 (2012)
80. Sipal, V., Allen, B., Edwards, D., Honary, B.: Twenty years of ultrawideband: opportunities and challenges. IET Commun. **6**(10), 1147 (2012)
81. Jusoh, M., et al.: A MIMO antenna design challenges for UWB application. Prog. Electromagnet. Res. B **36**, 357–371 (2012)
82. Adamiuk, G., Wiesbeck, W., Zwick, T.: Multi-mode antenna feed for ultra wideband technology, In: IEEE Radio Wirel. Symp. vol. 2009, pp. 578–581 (2009)
83. Tang, M.C., Wang, H., Deng, T., Ziolkowski, R.W.: Compact planar ultrawideband antennas with continuously tunable, independent band-notched filters. IEEE Trans. Antennas Propag. **64**(8), 3292–3301 (2016)
84. Tang, M.C., Shi, T., Ziolkowski, R.W.: Planar ultrawideband antennas with improved realized gain performance. IEEE Trans. Antennas Propag. **64**, 61–69 (2016)
85. Grosswindhager, B., et al.: Poster: Switchable directional antenna system for UWB-based internet of things applications. In: Proceedings of the 14th EWSN Conference (2017)
86. Bhatia, D., Kumar, D.M., Sharma, A.: A beam scanning UWB antenna system for wireless applications. Intern. J. Electron. Eng. **3**(1) (2011)
87. Mottola, L., et al.: Electronically-switched directional antennas for wireless sensor networks: a full-stack evaluation. In: IEEE International Conference on Sensing, Communication and Networking (2013)

88. Catarinucci, L., Guglielmi, S., Patrono, L., Tarricone, L.: Switched-beam antenna for wireless sensor network nodes. Prog. Electrom. Res. C **39**, 193–207 (2013)
89. Demirkol, I., Ersoy, C., Alagoz, F., et al.: MAC protocols for wireless sensor networks: a survey. IEEE Commun. Mag. **44**(4), 115–121 (2006)
90. Huang, P., Xiao, L., Soltani, S., et al.: The evolution of MAC protocols in wireless sensor networks: a survey. IEEE Commun. Surv. Tutor. **15**(1), 101–120 (2013)
91. Ye, W., Heidemann, J., Estrin, D.: An energy-efficient MAC protocol for wireless sensor networks, In: Joint Conference of the IEEE Computer and Communications Societies (2002)
92. Van Dam, T., Langendoen, K.: An adaptive energy-efficient MAC protocol for wireless sensor networks. In: Proceedings of the 1st SenSys Conference (2003)
93. Buettner, M., Yee, G.V., Anderson, E., Han, R.: X-MAC: a short preamble MAC protocol for duty-cycled wireless sensor networks. In: Proceedings of the 4th SenSys Conference (2006)
94. Schuß, M., Boano, C.A., Weber, M., Römer, K.: A competition to push the dependability of low-power wireless protocols to the edge. In: Proceedings of the 14th EWSN Conference (2017)
95. Burri, N., von Rickenbach, P., Wattenhofer, R.: Dozer: Ultra-low power data gathering in sensor networks. In: Proceedings of the 6th IPSN Conference (2007)
96. Bober, W., Bleakley, C.J.: Bailighpulse: a low duty cycle data gathering protocol for mostly-off wireless sensor networks. Comput. Netw. **69**, 51–65 (2014)
97. Boano, C.A., Römer, K.: External Radio Interference. In: Radio Link Quality Estimation in Low-Power Wireless Networks. Springer International Publishing (2013)
98. Wu, Y., Stankovic, J.A., He, T., Lin, S.: Realistic and efficient multi-channel communications in wireless sensor networks. In: Proceedings of the 27th IEEE INFOCOM Conference (2008)
99. Kim, Y., Shin, H., Cha, H.: Y-MAC: An energy-efficient multi-channel MAC protocol for dense wireless sensor networks. In: Proceedings of the 7th IEEE IPSN Conference (2008)
100. Rajendran, V., et al.: Energy-efficient, collision-free medium access control for wireless sensor networks. Wirel. Netw. **12**(1), 63–78 (2006)
101. van Hoesel, L., Havinga, P.: A lightweight medium access protocol (LMAC) for wireless sensor networks: reducing preamble transmissions and transceiver state switches. In: Proceedings of the 1^{st}, International Workshop on Networked Sensing Systems (INSS) (2004)
102. IEEE Standard for Local and metropolitan area networks. Part 15.4. Amendment 1: MAC sublayer. IEEE Computer Society Std. 802.15.4e (2012)
103. Abramson, N.: The ALOHA system: Another alternative for computer communications. In: Proceedings of the Fall Joint Computer Conference (1970)
104. Kleinrock, L., Tobagi, F.: Packet switching in radio channels: Part i-carrier sense multiple-access modes and their throughput-delay characteristics. IEEE Trans. Commun. **23**(12), 1400–1416 (1975)
105. Tobagi, F., Kleinrock, L.: Packet switching in radio channels: Part ii-the hidden terminal problem in carrier sense multiple-access and the busy-tone solution. IEEE Trans. Commun. **23**(12), 1417–1433 (1975)
106. Karn, P.: MACA-a new channel access method for packet radio. In: ARRL/CRRL Amateur radio 9th Computer Networking Conference, vol. 140 (1990)
107. IEEE Standard for Low-Rate Wireless Networks. IEEE Computer Society Standards, 802.15.4 (2015)
108. Zhou, G., Stankovic, J.A., Son, S.H.: Crowded spectrum in wireless sensor networks. In: Proceedings of the 3rd Workshop on Embedded Networked Sensors (EmNets) (2006)
109. Petrova, M., et al.: Interference measurements on performance degradation between colocated IEEE 802.11g/n and IEEE 802.15.4 networks. In: Proceedings of the 6th ICN Conference (2007)
110. IEEE Standard for Local and metropolitan area networks. Part 15.4. Amendment 1: Add Alternate PHYs. IEEE Standard 802.15.4a (2007)
111. Zhang, J., Orlik, P.V., Sahinoglu, Z., Molisch, A.F., Kinney, P.: UWB systems for wireless sensor networks. Proc. IEEE **97**(2), 313–331 (2009)

112. Catherwood, P.A., Scanlon, W.G.: Ultra-wideband communications-an idea whose time has still yet to come? IEEE Antennas Propag. Mag. **57**(2) (2015)
113. DW1000 Datasheet. Version 2.09. DecaWave Ltd. (2016)
114. Kempke, B., et al.: Surepoint: Exploiting ultra wideband flooding and diversity to provide robust, scalable, high-fidelity indoor localization. In: Proceedings of the 14th SenSys Conference (2016)
115. Conti, A., Dardari, D., Win, M.Z.: Experimental results on cooperative UWB based positioning systems. In: 2008 IEEE International Conference on Ultra-Wideband (2008)
116. Chehri, A., Fortier, P., Tardif, P.M.: UWB-based sensor networks for localization in mining environments. Ad Hoc Netw. **7**(5), 987–1000 (2009)
117. Chipara, O., Lu, C., Bailey, T.C. Roman, G.-C.: Reliable clinical monitoring using wireless sensor networks: Experiences in a step-down hospital unit. In: Proceedings of the 8th SenSys Conference, ser. SenSys '10 (2010)
118. Angelopoulos, C.M., et al.: A smart system for garden watering using wireless sensor networks. In: Proceedings of the 9th ACM Symp. on Mobility Management and Wireless Access (2011)
119. Shen, X., Zhuang, W., Jiang, H., Cai, J.: Medium access control in ultra-wideband wireless networks. IEEE Trans. Veh. Technol. **54**(5), 1663–1677 (2005)
120. Sousa, E.S., Silvester, J.A.: Spreading code protocols for distributed spread-spectrum packet radio networks. IEEE Trans. Commun. **36**(3), 272–281 (1988)
121. Karapistoli, E., et al.: MAC protocols for ultra-wideband ad hoc and sensor networking: A survey. In: 4th International Congress on Ultra Modern Telecommunications and Control Systems (2012)
122. Liang, C.-J.M., Priyantha, N.B., Liu, J., Terzis, A.: Surviving Wi-Fi interference in low power ZigBee networks. In: Proceedings of the 8th SenSys Conference (2010)
123. Rao, V.P., Marandin, D.: Adaptive backoff exponent algorithm for Zigbee (IEEE 802.15.4). In:International Conference on Next Generation Wired/Wireless Networking (2006)
124. Draves, R., Padhye, J., Zill, B.: Routing in multi-radio, multi-hop wireless mesh networks. In: Proceedings of the 10th International Conference on Mobile Computing and Networking (2004)
125. Beutel, J.: Fast-prototyping using the btnode platform. In: Proceedings of the Design Automation & Test in Europe Conference, vol. 1 (2006)
126. Lin, S., Zhang, J., Zhou, G., Gu, L., Stankovic, J.A., He, T.: ATPC: Adaptive transmission power control for wireless sensor networks. In: Proceedings of the 4th SenSys Conference (2006)
127. Cuomo, F., Martello, C., Baiocchi, A., et al.: Radio resource sharing for ad hoc networking with UWB. IEEE J. Sel. Areas Commun. **20**(9), 1722–1732 (2002)
128. DecaWave Ltd.: APR001 Part2 Application Note. Non line of sight operation and optimizations to improve performance in DW1000 based systems, version 1.4 (2014)
129. Watteyne, T., Lanzisera, S., Mehta, A., et al.: Mitigating multipath fading through channel hopping in wireless sensor networks. In: Proceedings of the International Conference on Communication (2010)
130. Song, J., Han, S., Mok, A., et al.: WirelessHART: Applying wireless technology in real-time industrial process control. In: Real-Time & Embedded Technology and Applications Symposium (2008)
131. ANT Alliance. https://www.thisisant.com/
132. Du, P., Roussos, G.: Adaptive time slotted channel hopping for wireless sensor networks. In: Proceedings of the 4th Computer Science and Electronic Engineering Conference (2012)
133. Hauer, J.-H., Handziski, V., Wolisz, A.: Experimental study of the impact of WLAN interference on IEEE 802.15.4 body area networks. In: Proceedings of the 6^{th} EWSN Conference (2009)
134. Musaloiu-E, R., Terzis, A.: Minimising the effect of WiFi interference in 802.15.4 wireless sensor networks. Int. J. Sens. Netw. **3**(1), 43–54 (2007)

135. Zúñiga, M.A., Irzynska, I., Hauer, J.-H., et al.: Link quality ranking: Getting the best out of unreliable links. In: Proceedings of the 7^{th} DCOSS Conference (2011)
136. Hermans, F., et al.: Light-weight approach to online detection and classification of interference in 802.15.4-based sensor networks. In: Proceedings of the 3rd CONET Workshop (2012)
137. Boers, N.M., et al.: Sampling and classifying interference patterns in a wireless sensor network. ACM Trans. Sens. Netw. **9**(1), 2:1–2:19 (2012)
138. Kerkez, B., Watteyne, T., Magliocco, M., et al.: Feasibility analysis of controller design for adaptive channel hopping. In: Proceedings of the 4^{th} Valuetools Conference (2009)
139. Xu, R., Shi, G., Luo, J., Zhao, Z., Shu, Y.: MuZi: Multi-channel ZigBee networks for avoiding WiFi interference. In: Proceedings of the 4^{th} CPSCOM Conference (2011)
140. Varshney, A., et al.: Directional transmissions and receptions for high-throughput bulk forwarding in wireless sensor networks. In: Proceedings of the 13th SenSys Conference (2015)
141. Giorgetti, G., Cidronali, A., Gupta, S.K.S., Manes, G.: Exploiting low-cost directional antennas in 2.4 GHz IEEE 802.15.4 wireless sensor networks. In: Proceedings of the EuWiT (2007)
142. Michalopoulou, A., Koxias, E., Lazarakis, F., et al.: Investigation of directional antennas effect on energy efficiency and reliability of the IEEE 802.15.4 standard in outdoor wireless sensor networks. In: Proceedings of the 15th MMS Symposium (2015)
143. Dai, H.-N., Ng, K.-W., Li, M., Wu, M.-Y.: An overview of using directional antennas in wireless networks. Int. J. Commun. Syst. **26**(4), 413–448 (2013)
144. Nasipuri, A., et al.: A MAC protocol for mobile ad hoc networks using directional antennas. In: Proceedings of the IEEE Wireless Communication and Networking Conference, vol. 3 (2000)
145. Radunovic, B., Le Boudec, J.-Y.: Optimal power control, scheduling, and routing in UWB networks. IEEE J. Sel. Areas in Commun. **22**(7), 1252–1270 (2004)
146. Qi, Y., et al.: Clear channel assessment (CCA) with multiplexed preamble symbols for impulse ultra-wideband (UWB) communications. In: IEEE International Conference on UWB (2006)
147. Lazik, P., Rajagopal, N., Shih, O., Sinopoli, B., Rowe, A.: ALPS: A bluetooth and ultrasound platform for mapping and localization. In: Proceedings of the 13th SenSys Conference, ser. SenSys '15 (2015)
148. Gezici, S., et al.: Localization via ultra-wideband radios: a look at positioning aspects for future sensor networks. IEEE Signal Process. Mag. **22**(4), 70–84 (2005)
149. Alcock, P., Brown, J., Roedig, U.: Implementation and evaluation of combined positioning and communication. In: Real-World Wireless Sensor Networks, pp. 126–137 (2010)
150. Cheong, P., Oppermann, I.: An energy-efficient positioning-enabled MAC protocol (PMAC) for UWB sensor networks. In : IST Mobile and Wireless Communication Summit (2005)
151. Kulmer, J., Hinteregger, S., Großwindhager, B., Rath, M., Bakr, M., Leitinger, E., Witrisal, K.: Using decawave UWB transceivers for high-accuracy multipath-assisted indoor positioning. In: IEEE ICC 2017 Workshop on Advances in Network Localization and Navigation (ANLN) (2017)
152. Greiner, P., Grosinger, J., Schweighofer, J., Steffan, C., Wilfling, S., Holweg, G., Bösch, W.: A system on chip crystal-less wireless sub-GHz transmitter. IEEE Trans. Microw. Theory Tech. (2017)

Part III
Crowdsensing and Smart Cities

User Incentivization in Mobile Crowdsensing Systems

Constantinos Marios Angelopoulos, Sotiris Nikoletseas, Theofanis P. Raptis and José Rolim

Abstract In this chapter, we present basic design issues of mobile crowdsensing systems (MCS) and investigate some characteristic challenges. We define the basic components of an MCS (the task, the server and the crowd), investigate the functions describing/governing their interactions and identify three qualitatively different types of tasks. For a given type of task, and a finite budget, the server makes offers to the agents of the crowd based on some incentive policy. On the other hand, each agent that receives an offer decides whether it will undertake the task or not, based on the inferred cost (computed via a Cost function) and some join policy. In their policies, the crowd and the server take into account several aspects, such as the number and quality of participating agents, the progress of execution of the task and possible network effects, present in real-life systems. We evaluate the impact and the performance of selected characteristic policies, for both the crowd and the server, in terms of task execution, budget efficiency and workload balance of the crowd. Experimental findings demonstrate key performance features of the various policies and indicate that some policies are more effective in enabling the server to efficiently manage its budget while providing satisfactory incentives to the crowd and effectively executing the system tasks. Interestingly, incentive policies that take into account the current

C. Marios Angelopoulos
Bournemouth University, Bournemouth, UK
e-mail: mangelopoulos@bournemouth.ac.uk

S. Nikoletseas
Department of Computer Engineering and Informatics, University of Patras and Computer Technology Institute and Press Diophantus (CTI), Patras, Greece
e-mail: nikole@cti.gr

T. P. Raptis (✉)
Institute of Informatics and Telematics, National Research Council, Pisa, Italy
e-mail: theofanis.raptis@iit.cnr.it

J. Rolim
University of Geneva, Geneva, Switzerland
e-mail: jose.rolim@unige.ch

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_8

crowd participation achieve a better trade-off between task completion and budget expense.

1 Introduction

During the past years, high adoption rates of truly portable smart devices, such as smartphones and other wearable devices (e.g. smart watches [1] and glasses [2]), have shaped a new technological reality [3]. Nowadays we are capable of freely moving around carrying in our pockets technological artefacts with significant computational and communication resources, while exchanging large volumes of data among us. This ubiquitous presence of smart devices, that have the capability of being always connected from everywhere, offers an unprecedented ability of augmenting traditional computer networks and systems with crowdsourced resources; i.e. with smart devices provided by the public. In this context, recently a new paradigm has emerged for distributed sensing systems and applications.

Mobile crowdsensing systems (*MCS*) instead of relying on special-purpose distributed systems, like Wireless Sensor Networks, they exploit the embedded sensory capabilities of modern smartphones (and of other similar devices) in order to collaboratively perform data collection [4]. Collecting data in a distributed manner from a set of autonomous devices available in an area of interest is not a novel idea. In fact, several aspects of such systems, like efficiency, robustness, scalability, and network lifetime, have been extensively studied during the past years. Even the notion of unpredictable and highly diverse mobility in such systems is not novel. However, the envisioned mobile crowdsensing systems demonstrate several characteristic attributes that clearly distinguish them from well-studied sensing systems, like Wireless Sensor Networks.

First, each node of a MCS (a smartphone, a tablet or other devices) has significantly more computational capabilities than a corresponding node of a traditional sensing system such as a sensor mote. Second, typical sensor motes only support one type of wireless interface (e.g. IEEE 802.15.4), thus requiring a gateway to act as a liaison between the network and the rest of the world. On the contrary, modern smartphones (and several tablets) are equipped with three or more, qualitatively different types of wireless communication interfaces. Last but not least, a major difference between traditional sensing systems and the envisioned MCS is the human factor. In MCS, each sensing point is controlled by a person that has to consent in order for its device to participate in the system. This need for consent adds a high degree of unpredictability and unreliability and raises the need to design incentive mechanisms in order to engage the owners of the devices, while taking into account their individual preferences and behaviour.

Furthermore, significant challenges are posed with respect to the ownership of several smart devices. In fact, smart devices and gadgets such as smartphones and smart wearables raise significant challenges in terms of trust, security and privacy for their owners. This is a factor that can potentially hinder successful, long-term

engagement of the crowd, along with concerns regarding use of device resources (e.g. battery drainage or data usage charges). Finally, challenges are raised with respect to data ownership and the role of commercial parties such as hardware providers (Apple, Samsung, etc.), service providers (e.g. Google, Apple, etc.) and third-party application developers. Such issues are dealt with not only on a technical but also on a regulatory and legislative level. For instance, via data protection laws, such as Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Such initiatives clearly define the operation framework of platforms that collect, store and manage data and safeguard the interests of the general public by establishing concepts such as the 'Right to be Forgotten'. In this book chapter, we focus on the design aspects of MCS from a Computer Science perspective.

In this chapter, we present some key design issues of a mobile crowdsensing system and identify its main characterizing challenges, coming from our recent line of work (presented in [3–8]). We particularly focus on the line of research implemented in [5, 6], and we review the basic components of an MCS—the task, the server and the crowd—and investigate the functions that describe/govern their interactions.

In particular, motivated by real-life applications [7, 8], we first review and provide examples of three qualitatively different types of tasks; (a) those whose added utility is proportional to the size of the task, (b) those whose added utility is proportional to the progress of the task and (c) those whose added utility is reversely proportional to the progress of the task. Then, we define the crowd as a set of autonomous agents each of which follows its own join policy and is characterized by its own attributes, such as a quality indicator and a personal threshold, based on which the incentives provided are evaluated. The server abstracts a stakeholder that wishes to utilize the augmented sensory capabilities provided by the crowd in order to perform a task. A finite budget $\mathcal{B}$ is at the disposal of the server for providing incentives to the crowd. The budget abstracts either monetary incentives, access to premium services (such as more Internet bandwidth) or any other kind of incentive. The budget needs to be managed efficiently in order to yield as much payoff from the crowd as possible and the server does so via a utility function and an incentive policy.

With respect to the identified types of tasks, we evaluate and identify the most suitable incentive policy the server should follow. Our performance evaluation is conducted via selected metrics, such as the percentage of task completion, the overall spent budget and the corresponding trade-off between the two, the workload balance of the network and the achieved cumulative quality of the performed task.

2 Related Work

During the past few years, smartphones and other truly portable devices (such as tablets, smart watches [1] and smart glasses [2]) have evolved into sophisticated multi-sensory computing platforms. In [9], an overview is provided of the current

state of applications that are based on MCS systems. The main challenges recognized refer to resource limitations, such as available energy, bandwidth and computational power, privacy issues that may arise due to correlation of sensor data with individuals and the lack of a unifying architecture that would optimize the cross-application usage of sensors on a particular device or even on a set of correlated devices (e.g. if they are located in the same geographical area). In [10], the authors recognize the opportunity of fusing information from populations of privately held sensors as well as the corresponding limitations due to privacy issues. In this context, they describe the principles of community-based sensing and they propose corresponding methods that take into consideration the uncertain availability of the sensors, the context-sensitive value of sensor information and sensor owners' preferences about privacy and resource usage. The authors present efficient and well-characterized approximations of optimal sensing policies in the context of a road traffic monitoring application.

In more recent works, in [11] the authors use the notion of Participatory Sensing (PS) to describe such systems. They consider the problem of efficient data acquisition methods for multiple PS applications while taking into consideration issues such as resource constraints, user privacy, data reliability and uncontrolled mobility. They evaluate heuristic algorithms that seek to maximize the total *social welfare* via simulations that are based on mobility datasets consisted of both real-life and artificial data traces. In [12], the authors propose a utility-driven smartphone middleware for executing community-driven sensing tasks. The proposed middleware framework considers preferences of the user and resources available on the phone to tune the sensing strategy thus enabling the execution of tasks in an opportunistic and passive manner.

In [13] the sensing capabilities of smart devices are classified into three distinct categories; inertial sensors (such as accelerometers and gyroscopes), positioning and proximity sensors (like GPS and information correlated to wireless access points) and ambient environment sensors (e.g. cameras, microphones, magnetometers, etc.). Data coming from such sensors can be used in order to extract several types of features regarding physical activities, social interactions and the environment. The feature extraction is achieved by employing a variety of techniques including among others discriminative models, decision trees, fuzzy logic and Bayesian classifiers.

By taking advantage of these capabilities, several proof-of-concept applications have been developed in a variety of topics including transportation, health, environmental monitoring and other. For instance, VTrack [14], developed in MIT, is a system for travel time estimation using smartphone sensor data. By utilizing methods like the hidden-Markov model and sparse data interpolation to process the sensory data, the system is able to provide accurate location estimates and corresponding delays for delay-aware routing algorithms. In [15], a crowdsourced approach of detecting and localizing events in outdoor environments is presented. Each smartphone user simply has to point his device towards the direction of an event in order for the application to collect and report sensory data including accelerometer, compass, GPS and time. By combining data from multiple users, the application is capable of successfully localizing events taking place nearby. What is of great interest is the

fact that although each individual measurement may be inaccurate, the final precision of the application is proportional to the number of total measurements; in other words, there appears a *network effect*. In [16], the authors present a scalable Internet system, designed for continuous video collection from crowdsourced devices such as smartphones and Google Glasses [2]. By decentralizing the collection infrastructure using virtual machines, the system achieves scalability while also providing a privacy-preserving mechanism that automatically removes sensitive information from the videos. In [17], the authors present a crowdedness detection scheme for mobile crowdsensing applications; i.e. a duty cycle adaptation scheme that provides an estimation of how dense the neighbourhood of a smartphone is. Finally, [18] refers to more crowdsensing applications while also providing a survey on mobile phone sensing.

Few programming frameworks have also been introduced in an effort to facilitate the design and development of crowdsensing applications. In [19] the MEDUSA programming framework is introduced that is specifically designed to address the particular requirements of crowdsensing applications. By providing high-level abstractions of commonly used subtasks the description of a crowdsourcing task is reduced by two orders of magnitude, while at the same time a distributed runtime system coordinates the task execution between several smartphones and a cluster on the cloud. A second development framework is PRISM [20], that adopts a push model that enables timely and scalable application deployment while ensuring a good degree of privacy. It manages to do so by enabling the application developers to package their applications as executable binaries that are then automatically deployed to smartphones based on some specified predicates.

The role of social media has also been studied as a crowdsourcing platform. In [21], the authors investigate the advantages and disadvantages of crowdsourcing applications applied to disaster relief coordination. It also discusses several challenges that must be addressed to make crowdsourcing a useful tool that can effectively facilitate the relief progress in coordination, accuracy and security. A similar work on [22] studies the effect of employing social media and other crowd-driven platforms on the World Wide Web. Although also referring to crowdsourcing, these works do not relate to the specific paradigm of MCS. Albeit a crowdsourcing paradigm, MCS provision data collection via mobile and portable devices; therefore, an MCS system is affected by and takes advantage of the geographical distribution of people and the corresponding spatiotemporal dynamics. Also, most commonly, MCS provision the collection of ambient data, usually related to the physical surroundings of people via corresponding sensors. On the contrary, the aforementioned works examine the role of social media as enablers in information exchange or crowd collaboration.

Apart from specific applications and application development frameworks, significant effort has been made to define models for the crowdsensing paradigm. In [23], the authors consider two system models; namely the platform-centric and the user-centric models. By using game-theoretic analysis for the first and auction theory for the latter, corresponding incentive mechanisms are designed for each model. Although the provided incentive mechanisms are well designed (e.g. for the user-centric model the mechanisms are efficient, rational, profitable and truthful), however,

they are based on the assumption that the agents and the server have full information regarding the task allocation procedure (e.g. what is the total task to be executed, what is the total available budget, etc.). Also, in [24], the authors study incentive mechanisms for a mobile crowdsensing scheduling problem, where a mobile crowdsensing application owner announces a set of sensing tasks, then human users (carrying mobile devices) compete for the tasks based on their respective sensing costs and available time periods, and finally the owner schedules as well as pays the users to maximize its own sensing revenue under a certain budget. In contrast to the above works, we study online scenarios for crowdsensing systems; i.e. the server does not have complete knowledge on the system and in some cases, the policies followed by both the server and the agents are adjusted to the way the task allocation and execution evolve over time.

In [25], the authors investigate the problem of task pricing and scheduling on crowdsourcing markets. Trying to maximize the likelihood of a proposed task to be accepted for execution by the crowd, a survival analysis model is employed to provide an algorithm for determining the optimal reward for a crowdsourced task. Again, here the server is assumed to have access to full market information. Finally in [26], two mechanisms for validating the tasks performed in crowdsourcing platforms are studied in terms of cost and accuracy. The first mechanism decides whether the reported results are truthful based on a majority decision while the second one relies on a control group to perform the validation.

Table 1 Notation used

The crowd $\mathcal{C}$	
A_i	An individual agent of the crowd
N	Size of the crowd (total number of agents)
$N(t)$	Percentage of the crowd participated in task execution until time t
q_{A_i}	Quality indicator of agent A_i
m_{A_i}	Number of times A_i has already contributed in task execution
$thres_i$	Threshold of A_i regarding the evaluation of offers
c_{A_i}	Inferred cost to agent A_i for executing a task segment
The server $\mathcal{S}$	
$\mathcal{B}$	Initially available budget
$\mathcal{B}(t)$	Residual budget at time t
u_k	Expected utility gained by $\mathcal{S}$ from task execution
I_k	Incentive provided by $\mathcal{S}$ for executing task segment $\mathcal{T}_k$
The task $\mathcal{T}$	
$\mathcal{T}$	Task of size λ
$\mathcal{T}_k$	Task segment of size λ_k
K	Total number of task segments
$\lambda(t)$	Total size of task segments completed by time t

In contrast to the above works, we here study some key issues of a mobile crowdsensing system and identify its main characterizing challenges. We define the basic components of an MCS—the task, the server and the crowd—and investigate the functions that describe/govern their interactions. Our focus is not on developing a prototype system, rather than (with respect to the identified types of tasks), on evaluating and identifying the most suitable incentive policy the server should follow. We provide a well-designed theoretical model and several tasks, utilities, incentive and joint policies. We note, however, that it is difficult to use this model as generic framework so as to serve all possible MCS applications (Table 1).

3 The Model

We view *crowdsensing* as the practice of utilizing the embedded sensory capabilities of smart devices provided by a community in order to perform a *task*.

Definition 1 ([5, 6]) We define a *mobile crowdsensing system* as a *distributed* system consisting of:

1. *the crowd*; i.e. a set of devices inside an area of interest that are equipped with embedded sensory capabilities and are carried by people.
2. *the server*; i.e. a stakeholder that seeks to utilize the augmented computational, communication and sensory capabilities of the crowd in order to perform a *task*.
3. a set of functions governing the interactions among the crowd and the server (i.e. incentive mechanisms, join policies, etc.).

In the following, we refer to a *mobile crowdsensing system* comprising of a crowd $\mathcal{C}$ and a server $\mathcal{S}$. $\mathcal{C}$ consists of *agents* that abstract people carrying smart portable devices (one device per agent), while $\mathcal{S}$ abstracts a stakeholder that seeks to exploit the sensory capabilities offered by the crowd in order to perform a task $\mathcal{T}$. The process of executing $\mathcal{T}$ takes place in rounds. At the beginning of each round, the server publishes offers to the crowd consisting of *task segments* of $\mathcal{T}$ along with corresponding incentives. The agents evaluate the offers and either accept to execute the task segment being offered and receive the corresponding incentive or they reject the offer. If at the end of the round there are any task segments left unexecuted, the same process is repeated until either all task segments are executed or the entire budget available to $\mathcal{S}$ has been spent.

3.1 The Task

We define by $\mathcal{T}$ the task of total size λ that server $\mathcal{S}$ seeks to execute by exploiting the sensory capabilities of the crowd $\mathcal{C}$. Depending on the context, the size λ of $\mathcal{T}$ may refer either to processing effort (e.g. in FLOPs) or to the time interval (e.g. in

seconds) needed by the crowd in order to perform $\mathcal{T}$ (for example consider a target tracking application or an application monitoring an environmental attribute for a given amount of time).

Whether $\mathcal{S}$ has one or several tasks to be performed, without loss of generality we can assume that the server is able to break a given task $\mathcal{T}$ into several task segments $\mathcal{T}_k : k = \{1, 2, ..., K\}$ such that $\mathcal{T} = \bigcup_{k=\{1,2,...,K\}} \mathcal{T}_k$ and $\lambda \leq \sum_k \lambda_k$. This implies that the task segments could be overlapping over time. However, in this chapter, we consider the special case where $\mathcal{T}^{\lambda}$ is partitioned into equally sized, non-overlapping task segments. Finally, by $\lambda(t)$ we denote the cumulative size of task segments that have been executed by time t.

The server $\mathcal{S}$ tries to make the most out of the MCS, in terms of task execution, by efficiently managing the available budget $\mathcal{B}$. In fact, $\mathcal{S}$ will provide incentives to the agents by first evaluating the expected payoff gained by the execution of a task segment and second by offering a corresponding fraction of the budget to the agents.

Comment. At this point, we would like to note that, as stated before, the individual threshold $thres_i$ based on which each agent A_i evaluates the offers made is unknown to the server. Also, in this chapter, we do not investigate any strategies that the server could employ in order to infer $thres_i$ for each agent. Therefore, thresholds are unknown and unpredictable to the server and as such are considered random. Furthermore, we also consider that each task $\mathcal{T}$ is broken down to K equally sized and non-overlapping task segments. Therefore, we consider the server to allocate task segments and make the corresponding offers to the crowd by selecting agents uniformly at random.

The task utility. Depending on the type of the task $\mathcal{T}$, the server may have a different assessment of the expected payoff provided by the execution of a task segment. For instance, in one scenario it may suffice to receive live-streaming video regarding an event from only one agent; in this case, the expected utility from each consequent task segment gained from a second participating agent would be much less if not zero. On the other hand, in a localization scenario the more agents participate, the higher the accuracy of the tracking; here, the expected utility is proportional to the number of participating agents. In general, we consider that the expected utility of $\mathcal{S}$ received by the execution of task segment $\mathcal{T}_k$ is of the form $u_k = f(\lambda, \lambda_k, q_{A_i})$. Following, we identify three different qualities of utility functions and provide corresponding indicative task examples.

1. **Utility proportional to the task completion** ([5, 6]):

$$u_k = \frac{\lambda_k}{\lambda} \tag{1}$$

i.e. the expected utility gained for $\mathcal{S}$ is proportional to the size of the task segment to be executed. As an example consider an environmental monitoring application (e.g. monitoring background noise), where λ_k corresponds to the amount of time the crowd will be providing noise measurements. The longer the time interval

(corresponding to more task segments), the more information the server will collect.

2. **Utility proportional to the progress of the task** ([5, 6]):

$$u_k = \frac{(\lambda(t) + \lambda_k)^{\delta}}{\lambda} \tag{2}$$

where $0 < \delta < 1$, i.e. the expected utility gained for $\mathcal{S}$ is increasing over the overall task progress. As an example consider a video rendering application in which if one task segment is not executed, then the entire task $\mathcal{T}$ fails. In that case, as the spent budget on already executed task segments is increasing, the expected utility for the remaining task segments is also increasing. For instance, consider the case where the very last task segment fails to be executed; then the server will have spent almost the entire budget while the entire task will also have failed.

3. **Utility reversely proportional to the progress of the task** ([5, 6]):

$$u_k = \frac{\lambda_k}{\lambda(t) + d} \tag{3}$$

where d positive constant and $\lambda(t)$ the percentage of task completion by time t, i.e. the expected utility for $\mathcal{S}$ decreases over the progress of the execution of the overall task $\mathcal{T}$. In other words, as more and more task segments are executed, the expected utility from the remaining task segments is less. As an example consider a target tracking application; initially, as the first agents join the application the tracking accuracy is significantly improved, thus the expected utility is high. However, once the number of agents has reached a point that provides the desired tracking accuracy, the utility gained from any additional participating agents decreases since their participation does not provide additional information.

3.2 The Server

We define as server $\mathcal{S}$ a stakeholder that seeks to exploit the sensory and computational capabilities provided by the MCS in order to perform a task $\mathcal{T}$. A task could be, for example to monitor the background noise level for a given period of time or to remotely overlook an event by collecting live-streaming video. The server also has at its disposal a finite budget $\mathcal{B}$ that can freely manage in order to provide incentives to the agents of the crowd in order to participate in the execution of task $\mathcal{T}$. The nature of $\mathcal{B}$ can either be monetary or it may come in the form of a service, such as increased internet bandwidth allocation. We denote by $\mathcal{B}(t)$ the residual budget of the server at time t.

The incentive policy of the server. There are several strategies based on which the server $\mathcal{S}$ can manage the available budget $\mathcal{B}$. In general, we consider that the incentive

provided by $\mathcal{S}$ is of the form $I_k = f(u_k, N(t), \mathcal{B}(t))$. Following we identify four indicative incentive policies.

1. **Proportional incentive policy** ([5, 6]):

$$I_k = u_k \mathcal{B}(t) \tag{4}$$

Following this policy, the incentive allocated by $\mathcal{S}$ to each task segment is proportional to the expected utility and the current residual budget.

2. **Participation-aware incentive policy** ([5, 6]):

$$I_k = \frac{1}{c(N(t)+1)} u_k \mathcal{B}(t) \tag{5}$$

where c a positive constant. Following this policy, $\mathcal{S}$ initially provides high incentives in order to stimulate the crowd and achieve a minimum percentage of participating agents. Then $\mathcal{S}$ becomes more conservative, trying not to attract new agents but to sustain the already participating ones.

3. **Quality-aware incentive policy** ([5, 6]):

$$I_{A_i,k} = u_k \frac{q_{A_i}}{q_{max}} \mathcal{B}(t) \tag{6}$$

where q_{max} is the maximum quality that can be provided by a single agent of the crowd. Following this incentive policy, the incentive allocated by $\mathcal{S}$ to each task segment is proportional to the execution quality of the agent. This policy aims at attracting high-quality agents by offering higher amounts of incentive.

4. **Thrifty incentive policy** ([5, 6]):

$$I_k = u_k \left(\frac{\mathcal{B}(t)}{\mathcal{B}} \right)^{\epsilon} \mathcal{B}(t) \tag{7}$$

where ϵ a positive constant. This incentive policy, although not using additional crowd-based information (like q_{A_i} or $N(t)$), aims at a more restrained budget expenditure, by reinforcing the fraction of the residual and initial budget in the incentive computation. This means that the more the budget is spent throughout time, the less the quantity $\frac{\mathcal{B}(t)}{\mathcal{B}}$ becomes, a fact that results in a sharp drop of the eventual budget offered. It is designed for applications where the server utility acquisition requires high budget expenses.

3.3 The Crowd

We define as crowd $\mathcal{C}$ the set of devices $A_i : i \in \{1, 2, ..., N\}$ carried by agents inside an area of interest. The devices are characterized by some embedded sensory capabilities (e.g. accelerometers, gyroscopes, microphones, cameras, etc.) and are available to potentially undertake the execution of a task (or task segment) assigned by the server $\mathcal{S}$. The evaluation of the received offers is performed by each agent based on a characterizing threshold $thres_i$, which is unknown to the rest of the agents and $\mathcal{S}$. Each agent A_i is also characterized by a task execution quality indicator q_{A_i}, which depending on the context of the crowdsensing application may refer to computational power (e.g. FLOPs per second) or other application-specific attributes (e.g. camera resolution, quality of sound, etc.). Also, each agent is able to keep track of the number of times he has contributed to the execution of a task by maintaining a counter m_{A_i}. Finally, we denote by $N(t) \in [0, 1]$ the percentage of agents that have participated in the execution of a task at time t; i.e. number of participating agents over the total number of agents.

Once an agent A_i belonging to the crowd $\mathcal{C}$ has received an offer from the server $\mathcal{S}$ (i.e. a task segment to be executed with a corresponding incentive) first evaluates the cost that the task execution will infer (this cost may reflect energy dissipation, resource allocation over time, etc.) and then will make a decision on whether will it undertake the task segment or not.

The cost function of the agents. The execution of a task segment $\mathcal{T}_k$ by agent A_i, infers to the agent a cost computed by the agent's cost function. For a given $\mathcal{T}_k$, we consider the cost function to be of the form $c_{A_i} = f(\lambda_k)$; i.e. we consider the inferred to the agent cost to be proportional to the size of the allocated task segment. We identify the following cost function for the agents:

$$c_{A_i} = \alpha_i \lambda_k^{\beta_i} \tag{8}$$

where α_i, β_i are constants depending on each individual agent.

The join policy of the agents. Depending on the cost inferred by task $\mathcal{T}_k$ and the incentive I_k offered, the agent decides whether she will accept or decline the offer based on its join policy; i.e. a Boolean function $P_{A_i}(c_{A_i}, I_k, m_{A_i})$. Each agent is also individually characterized by a threshold $thres_i$ that is an independent variable and constitutes a measure based on which each agent evaluates the offers provided. $thres_i$ captures how willing the agent is to participate in the execution of the task and it has a varying impact according to the join policy of the agent. For instance, $thres_i$ has a decreasing impact when network effect phenomena are present and an increasing impact when the agent takes into account its past contributions in the task execution. We below identify three join policies of indicative qualities:

1. **Simple join policy** ([5, 6]):

$$P_{A_i,k} = \begin{cases} 1 \text{ if} \frac{I_k}{C_{A_i}} \geq thres_i \\ 0 \qquad \text{otherwise} \end{cases} \tag{9}$$

Agents following this policy simply compare the ratio between the incentive being offered over the expected inferred cost to their own threshold $thres_i$. If the ratio is higher than the threshold of the agent, then the agent accepts the offer, otherwise the offer is rejected.

2. **Join policy with network effect** ([5, 6]):

$$P_{A_i,k} = \begin{cases} 1 \text{ if } \frac{I_k}{C_{A_i}} \geq \frac{thres_i}{\eta N(t)} \\ 0 \qquad \text{otherwise} \end{cases} \tag{10}$$

where η is a constant. This policy captures network effect phenomena present in real-life systems according to which the more popular an application is the more willing people are to participate in it; e.g. social applications. In particular, additionally to the $thres_i$, the agent also takes into account the percentage of participating agents $N(t)$ when evaluating an offer.

3. **Join policy with memory** ([5, 6]):

$$P_{A_i,k} = \begin{cases} 1 \text{ if } \frac{I_k}{C_{A_i}} \geq thres_i \cdot m_{A_i}^{\gamma} \\ 0 \qquad \text{otherwise} \end{cases} \tag{11}$$

where γ is a constant. This policy captures the growing unwillingness of agents that are frequently chosen by $\mathcal{S}$. This unwillingness is due to the intense usage or high dissipation rate of their resources. In particular, each agent also takes into account the number of times m_{A_i} it has already participated in the application when evaluating an offer.

4 Performance Evaluation

4.1 Experimental Setup and Metrics

The experiments were conducted in Matlab R2013b with the following setup. We consider a crowd instance of $N = 100$ agents which are divided into three types according to their threshold values. There are the willing agents which have very low threshold value, the unwilling agents which have a very high threshold value and the agents in between whose threshold value is moderate with $200 < thres_i < 6000$. The agents also have varying agent qualities, with $0 < q_{A_i} < 10$. We consider $K = 1000$ equal-sized, non-overlapping task segments and the initial budget is set to $\mathcal{B} = 1000$ units. We set the constants values to $\eta = 100$ and $\gamma = 2$. In order to achieve statistical smoothness, we applied several times the deployment of nodes in the network and repeated each experiment 100 times. The statistical analysis of the findings (the median, lower and upper quartiles and outliers of the samples) demonstrate very

high concentration around the mean, so in the following figures, we only depict average values. Our evaluation is focused on the following performance metrics:

Percentage of task $\mathcal{T}$ completion over time. In particular, we will evaluate the utility functions and incentive policies of the server in terms of task completion over several types of crowd (where by the term "type" we refer to the join policy the agents follow). In other words, given the budget $\mathcal{B}$, we evaluate the expected number of executed task segments achieved by each configuration.

Percentage of residual budget $\mathcal{B}(t)$ over time. With this metric, we will evaluate the utility functions and incentive policies of the server in terms of the expected spending rate of the budget achieved over several types of crowd.

Expenditure efficiency (task completion over budget spent). With this metric, we wish to investigate the trade-off between budget spent and task segment execution rate. In other words, we wish to measure the efficiency of each utility function and incentive policy.

Workload balance of the crowd. With this metric, we wish to investigate how are task segments distributed over the agents. Although the task segment allocation is performed uniformly at random by the server, however different policies may favour different agents. For instance, consider a policy according to which the server offers very small incentives; in this case, only the very willing agents would accept the offers made and therefore the task execution would be imbalanced.

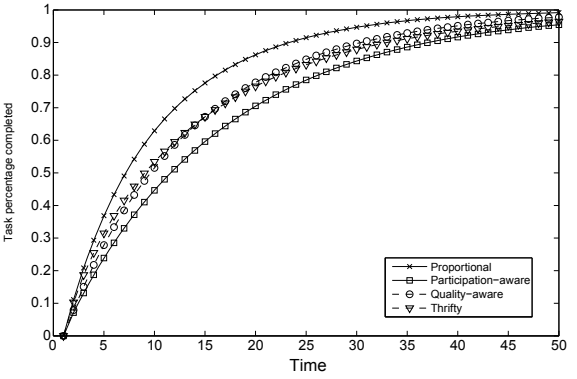
Task quality achieved. In particular, we will investigate the expected task quality achieved by the server; that is the accumulated quality of the executed task segments $\sum q_{A_i} \lambda_i$.

4.2 Incentive Policies' Performance

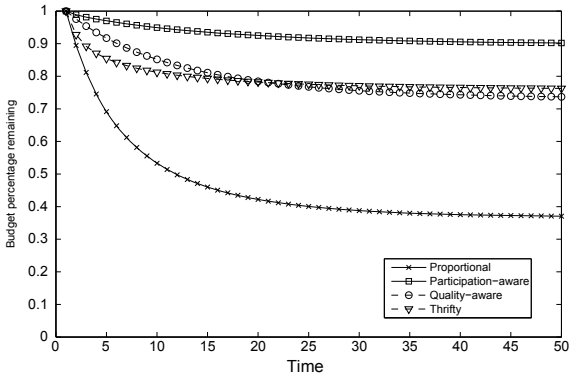
The performance of the four incentive policies under various utility functions, in terms of task completion and residual budget over time as well as expenditure efficiency, is depicted in Figs. 1, 2 and 3 respectively for the simple join policy, in Figs. 4, 5 and 6 for the join policy with network effect and in Figs. 7, 8 and 9 for the join policy with memory. For the simple join policy, as shown in Fig. 1a, all incentive policies persuade the agents to gradually complete the task segments, with almost the same rate. However, it is clear (Fig. 1b) that, for the participation-aware incentive policy, this is achieved with much lower budget overhead. This fact also becomes clear, after a closer look in Fig. 1c. By examinating the corresponding Figs. 4 and 7, we end up in the same conclusions for the join policies with network effect and with memory.

The behaviour of the incentive policies under the decreasing utility function is shown in Figs. 2, 5 and 8 for each join policy. In this case, the task execution over time is following almost the exact same pattern for all incentive policies. However, the rate of the budget expenditure is even more sharp for the proportional incentive policy, than the case of the proportional utility function. Also, the budget is expended at a very fast rate at the beginning of the task assignment process, in contrast to the proportional

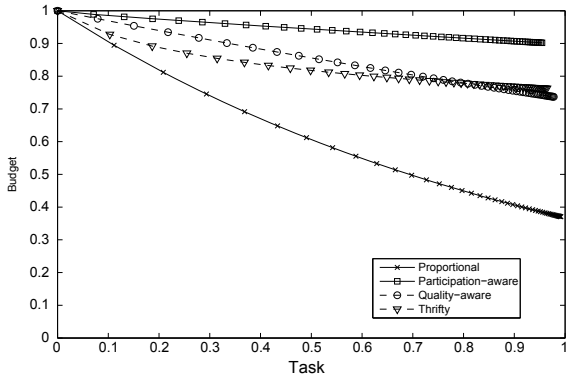
Fig. 1 Task and budget over time. Proportional utility function. Simple join policy



(a) Task percentage completed.

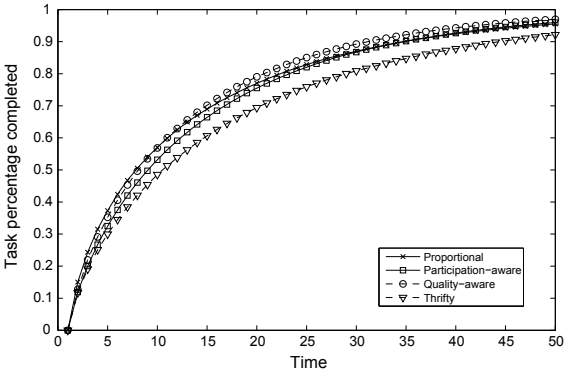


(b) Residual budget percentage.

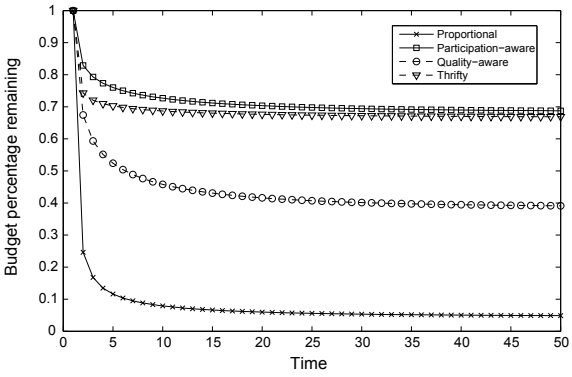


(c) Expenditure efficiency.

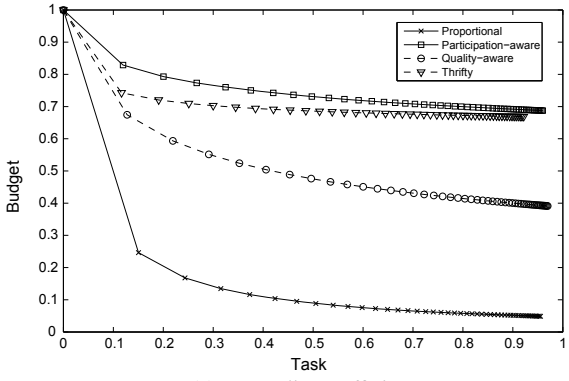
Fig. 2 Task and budget over time. Decreasing utility function. Simple join policy



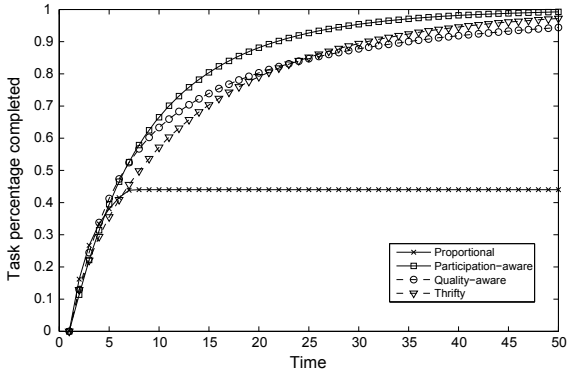
(a) Task percentage completed.



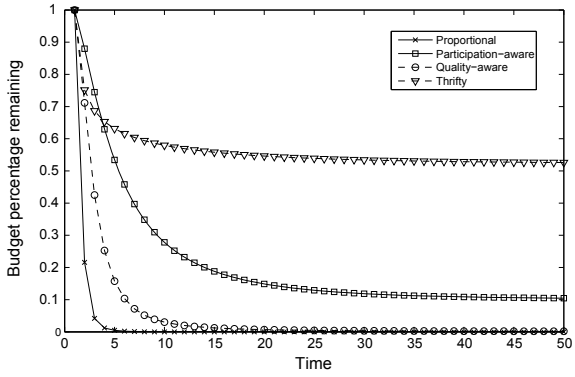
(b) Residual budget percentage.



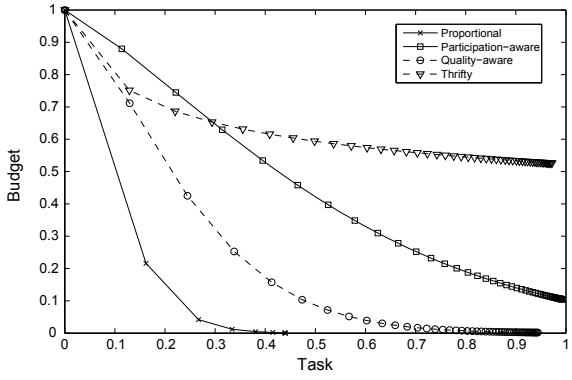
(c) Expenditure efficiency.



(a) Task percentage completed.



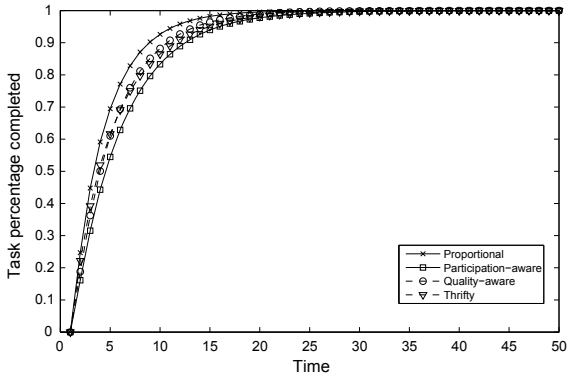
(b) Residual budget percentage.



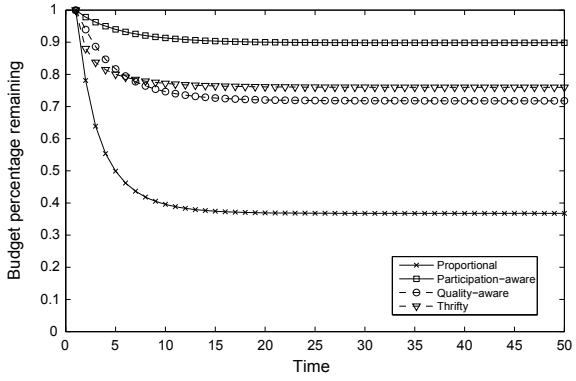
(c) Expenditure efficiency.

Fig. 3 Task and budget over time. Increasing utility function. Simple join policy

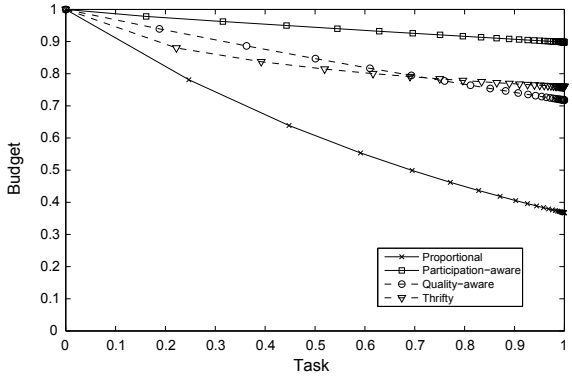
Fig. 4 Task and budget over time. Proportional utility function. Join policy with network effect



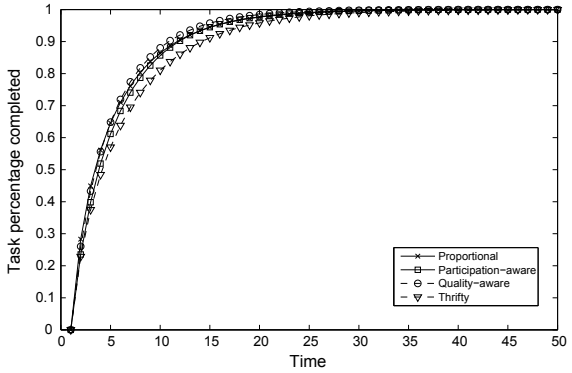
(a) Task percentage completed.



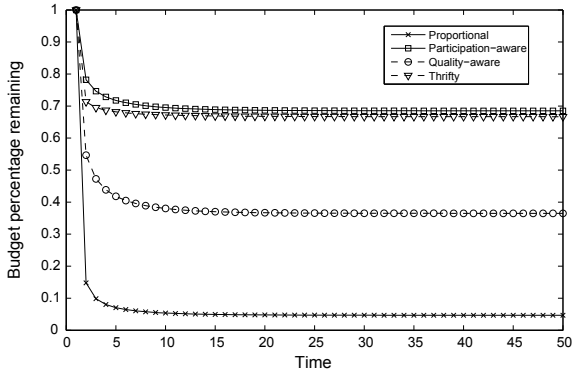
(b) Residual budget percentage.



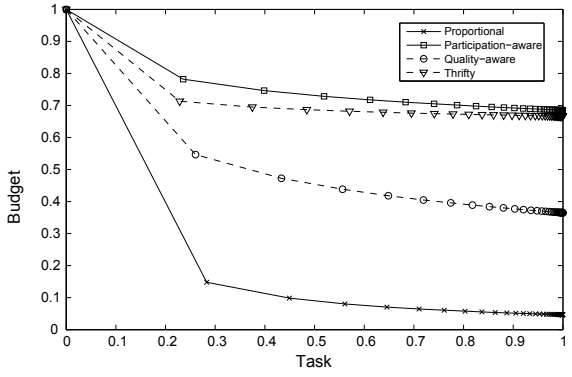
(c) Expenditure efficiency.



(a) Task percentage completed.

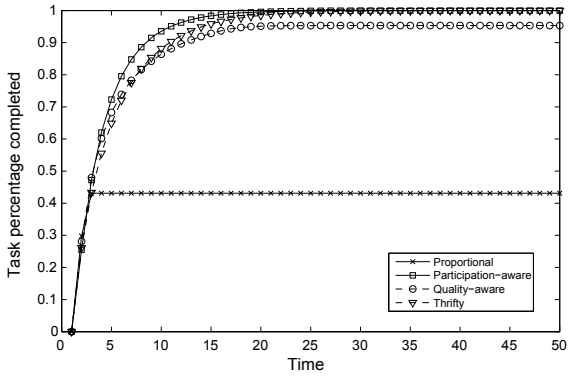


(b) Residual budget percentage.

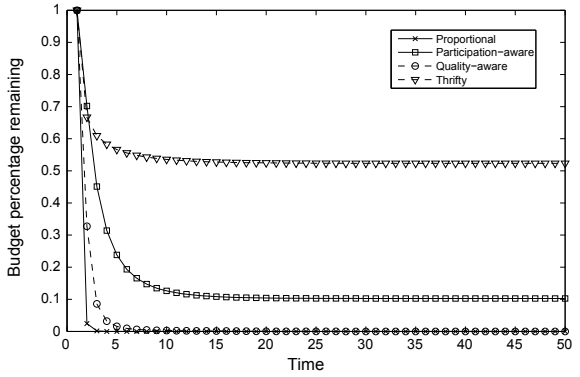


(c) Expenditure efficiency.

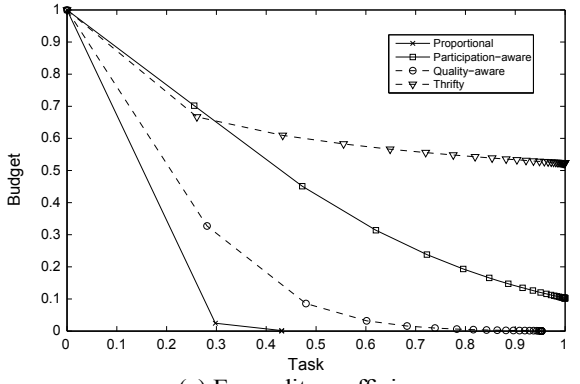
Fig. 5 Task and budget over time. Decreasing utility function. Join policy with network effect



(a) Task percentage completed.

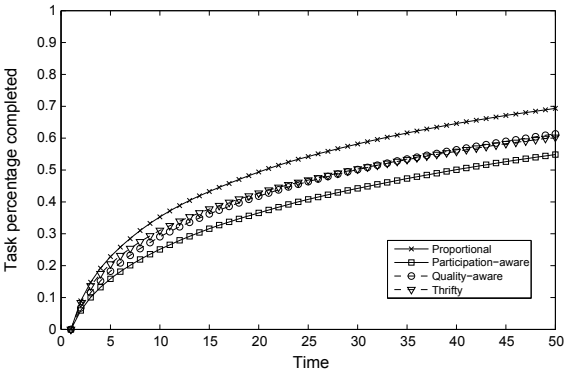


(b) Residual budget percentage.

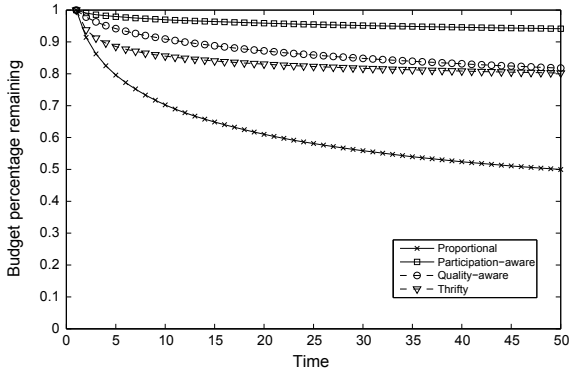


(c) Expenditure efficiency.

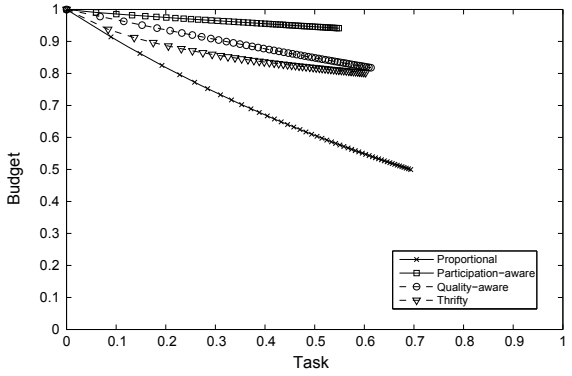
Fig. 6 Task and budget over time. Increasing utility function. Join policy with network effect



(a) Task percentage completed.

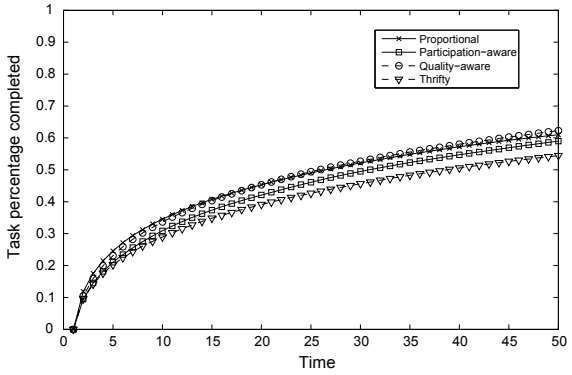


(b) Residual budget percentage.

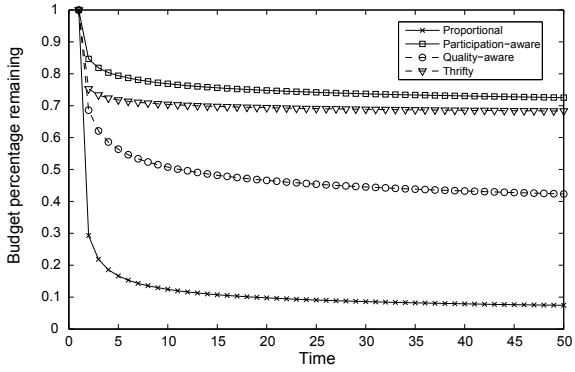


(c) Expenditure efficiency.

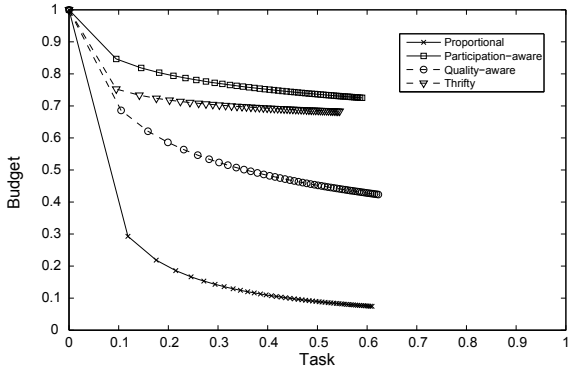
Fig. 7 Task and budget over time. Proportional utility function. Join policy with memory



(a) Task percentage completed.

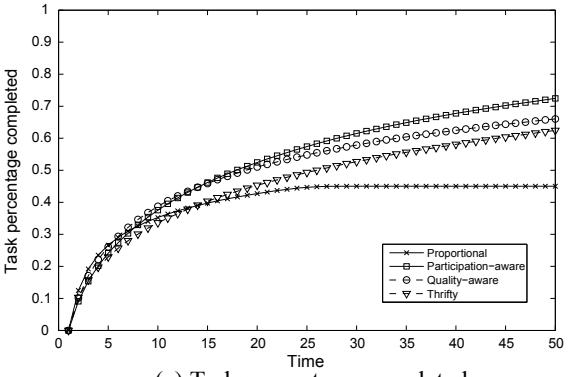


(b) Residual budget percentage.

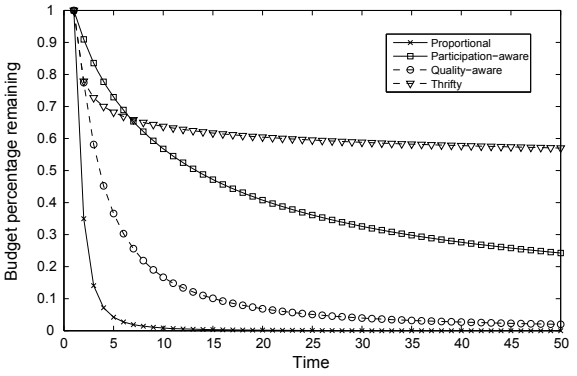


(c) Expenditure efficiency.

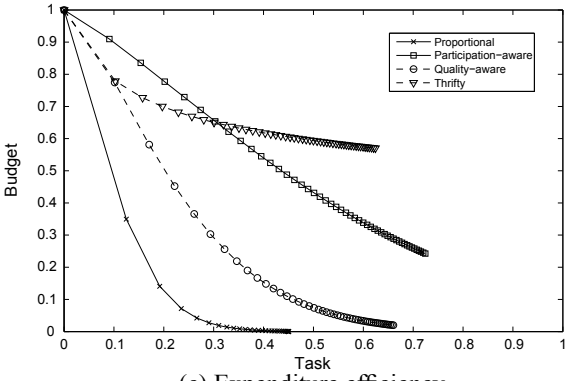
Fig. 8 Task and budget over time. Decreasing utility function. Join policy with memory



(a) Task percentage completed.



(b) Residual budget percentage.



(c) Expenditure efficiency.

Fig. 9 Task and budget over time. Increasing utility function. Join policy with memory

utility function case. This is explained by the nature of the decreasing utility function, where once the number of agents has reached a point that provides the desired accuracy, the utility gained from any additional participating agents decreases, since their participation does not provide additional information. This fact results in a sharp decrease of the offered amount of incentive.

Regarding the increasing utility function, it is clear that all incentive policies outperform the proportional one, in terms of task completion over time, budget remaining over time and expenditure efficiency (Figs. 3, 6 and 9). Some notable differences from the previous configurations are the task completion rate of the proportional incentive policy, which in this case halts due to expenditure inefficiency, and the budget expenditure of the participation-aware incentive policy which is more sharp. In this case, the thrifty incentive policy, which succeeds in budget management, while completing a high percentage of task segments, achieves a high overall expenditure efficiency, compared to other incentive policies.

Overall, the participation-aware incentive policy, which takes into account the current crowd participation, achieves a very good trade-off between task completion and budget expense via a network effect introduced.

Comment. The join policy modelling is validated by the simulations, if we observe carefully the three different Figures' sets that correspond to each join policy (Figs. 1, 2, 3, 4, 5, 6, 7, 8 and 9). When the agents are joining the experiment by taking into account the network effect, the task completion rate is higher than the simple join case, since the more the agents which join, the more the overall joining eagerness becomes. On the contrary, when the agents are using memory, the task completion rate decreases, because of their growing unwillingness to frequently participate in the experiment.

4.3 Workload Balance and Overall Task Quality

The cumulative task quality for incentive policies and utility functions combinations is shown in Table 2. The best task quality for each combination is marked as bold. In general, the quality-aware incentive policy achieves the best overall task quality. This fact is explained by its incentive distribution strategy, and more specifically by the proportional to the agent's quality incentive allocation. The lowest quality is achieved by the combination of the proportional incentive policy with the increasing utility function. Note that when the agents follow the join policy with memory, the overall quality achieved is lower compared to the two other join cases. This can be explained by Figs. 7a, 8a and 9a, from which we conclude that in this case the task completion rate is lower.

The average crowd workload and its standard deviation for the simple join function and the join function with memory are shown in Figs. 10a, b, respectively. Each point represents a combination of an incentive policy and a utility function. The perfectly balanced workload is marked with a straight line with zero deviation on $K/N = 10$ task segments per agent. Numbers 1–3 stand for the corresponding utility functions

Table 2 Quality achieved

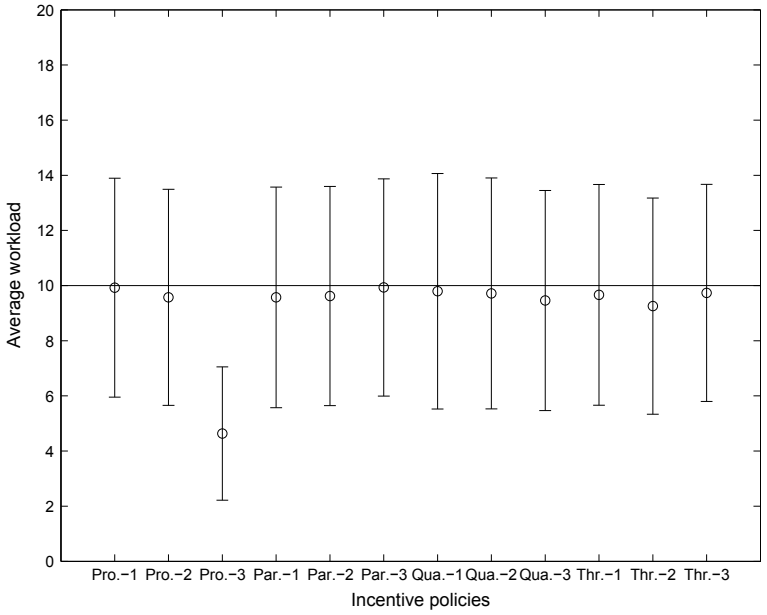
Utility	Incentive			
	Proportional	Participation-aware	Quality-aware	Thrifty
Simple join policy				
Proportional	6,279	6,099	**6,937**	6,167
Decreasing	6,078	6,111	**6,896**	5,898
Increasing	2,923	6,273	**6,602**	6,155
Join policy with network effect				
Proportional	5,801	5,631	**6,510**	5,649
Decreasing	5,622	5,633	**6,464**	5,465
Increasing	2,699	5,775	**6,081**	5,693
Join policy with memory				
Proportional	4.004	3.384	**4.051**	3.707
Decreasing	3.754	3.636	**4.109**	3.349
Increasing	2.785	**4.470**	4.341	3.842

(proportional, decreasing, increasing). When the crowd applies a simple join policy, the average workload among the agents is more balanced in almost all cases (except the case of proportional incentive and proportional utility). On the other hand, when the crowd applies a join policy with memory, the average workload is less balanced, with each agent being unwilling to overtake proposed tasks after a number of times having participated. This fact can also be observed in Figs. 7a, 8b and 9a, in which the behaviour of the crowd is more visible.

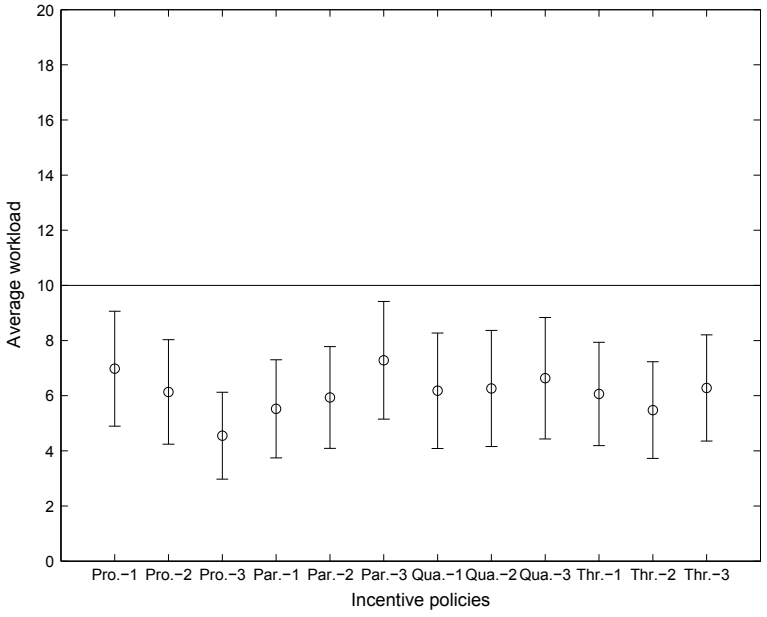
4.4 Utility Function and Incentive Policy Impact

The impact of the three utility functions, in terms of task completion and residual budget over time as well as expenditure efficiency, varies when using different incentive policies. Regarding the proportional incentive policy, the proportional and the decreasing utility functions are the most progressive in terms of task completion (Figs. 1a and 2a) whereas the increasing utility function is the least efficient (Fig. 3a). In contrary, for the participation-aware incentive policy, the application of an increasing utility function outperforms other utility functions application in terms of task completion over time. In spite of this efficiency, in this case, the participation-aware policy is not as cost efficient as the thrifty incentive policy.

Different types of tasks necessitate different approaches in terms of incentive policy applied. For the tasks corresponding to the proportional utility function, in which the gains of the server are proportional to the interval of the task completed, we observe that the best-applied incentive policy is the proportional one (Figs. 1a,



(a) Simple join policy.



(b) Join policy with memory.

Fig. 10 Workload balance of the crowd

4a and 7). As an example, consider an environmental monitoring application (e.g. monitoring background noise), where λ_k corresponds to the amount of time an agent is providing noise measurements. The longer the time interval (corresponding to more task segments), the more information the server will collect. For this reason, when the proportional incentive policy is applied, a steady time measurement task segment completion is maintained.

On the contrary, the proportional incentive policy is not suitable for tasks corresponding to the increasing utility function. In this type of tasks, since the value of each additional segment is even higher, other types of incentive policies should be considered (3a, 6a and 9a). As an example consider a video rendering application in which if one task segment is not executed, then the entire task T fails. In that case, as the spent budget on already executed task segments is increasing, the expected utility for the remaining task segments is also increasing. For instance, consider the case where the very last task segment fails to be executed; then the server will have spent almost the entire budget while the entire task will also have failed. In applications like video rendering, the participation-aware incentive policy performs good, because it maintains a pool of dedicated users that are frequently rewarded and thus are more keen on completing the corresponding task segments even at the later stages of the task.

As for the tasks corresponding to the decreasing utility function, the factor of quality is dominating the incentivization process, since after some time, the percentage of task completion is becoming less important than quality achieved. Consequently, the quality-aware incentive policy performs well (Figs. 2a, 5a and 8a). As an example consider a target tracking application; initially, as the first agents join the application the tracking accuracy is significantly improved, thus the expected utility is high. However, once the number of agents has reached a point that provides the desired tracking accuracy, the utility gained from any additional participating agents decreases since their participation does not provide additional information. The quality-aware incentive policy then attracts only a few high-quality agents by offering them higher amounts of incentive.

5 Conclusions and Future Work

In this work, we identified some key design issues of a mobile crowdsensing system and investigated some important characterizing challenges. We defined the basic components of an MCS, the crowd, the server and the task, and investigated the functions describing/governing their interactions. We evaluated the impact and the performance of selected characteristic policies, for both the crowd and the server, in terms of task execution, budget efficiency and workload balance of the crowd. Experimental findings indicate that some policies are more effective in enabling the server to efficiently manage its budget while providing satisfactory incentives to the crowd and affectively executing the system tasks.

For future research, we plan to further finetune the proposed model and investigate other cases of MCSs that are also characterized by realistic features, such as overlapping and non-equal task segments, varying crowd sizes and qualities over time, agent ability to entering or leaving a crowd, etc. We also plan to adopt business models in our research, both in the utility function design and in the incentive/join mechanisms.

References

1. Samsung. Samsung gear. http://www.samsung.com/global/microsite/galaxynote3-gear/spec.html
2. Google. Google glasses. https://support.google.com/glass/answer/3064128?hl=en&ref_topic=3063354
3. Angelopoulos, C.M., Filios, G., Nikoletseas, S., Raptis, T.P., Rolim, J., Veroutis, K., Ziegler, S.: Towards a holistic federation of secure crowd-enabled IoT facilities. In: Proceedings of IEEE ICC (2015)
4. Nikoletseas, S., Rapti, M., Raptis, T.P., Veroutis, K.: Decentralizing and adding portability to an IoT test-bed through smartphones. In: Proceedings of IEEE DCOSS (2014)
5. Angelopoulos, C.M., Nikoletseas, S., Raptis, T.P., Rolim, J.: Characteristic utilities, join policies and efficient incentives in mobile crowdsensing systems. In: Proceedings of IFIP Wireless Days (2014)
6. Angelopoulos, C.M., Nikoletseas, S., Raptis, T.P., Rolim, J.: Design and evaluation of characteristic incentive mechanisms in mobile crowdsensing systems. Simul. Model. Pract. Theory **55**, 95–106 (2015)
7. Angelopoulos, C.M., Evangelatos, O., Nikoletseas, S., Raptis, T.P., Rolim, J., Veroutis, K.: A user-enabled testbed architecture with mobile crowdsensing support for smart, green buildings. In: Proceedings of IEEE ICC (2015)
8. Fernandes, J., Krco, S., Rankov, A., Jokic, S., Nati, M., Loumis, N., Angelopoulos, C.M., Nikoletseas, S., Raptis, T.P., Ziegler, S.: IoT lab: towards co-design and IoT solution testing using the crowd. In: Proceedings of IEEE RIoT (2015)
9. Ganti, R.K., Ye, F., Lei, H.: Mobile crowdsensing: current state and future challenges. IEEE Commun. Mag. **49**, 32–39 (2011)
10. Krause, A., Horvitz, E., Kansal, A., Zhao, F.: Toward community sensing. In: International Conference on Information Processing in Sensor Networks, 2008. IPSN '08 (2008)
11. Riahi, M., Papaioannou, T.G., Trummer, I., Aberer, K.: Utility-driven data acquisition in participatory sensing. In: Proceedings of EDBT (2013)
12. Agarwal, V., Banerjee, N., Chakraborty, D., Mittal, S.: USense—a smartphone middleware for community sensing. In: Proceedings of MDM (2013)
13. Hoseini-Tabatabaei, S.A., Gluhak, A., Tafazolli, R.: A survey on smartphone-based systems for opportunistic user context recognition. ACM Comput. Surv. **45**(27), 1–51 (2013)
14. Thiagarajan, A., Ravindranath, L., Lacurts, K., Toledo, S., Eriksson, J., Madden, S., Balakrishnan, H., Chicago, U.I.: VTrack: accurate, energy-aware road traffic delay estimation using mobile phones (2009)
15. Ouyang, R.W., Srivastava, A., Prabahar, P., Roy Choudhury, P., Addicott, M., McClernon, F.J.: If you see something, swipe towards it: crowdsourced event localization using smartphones. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, ser. UbiComp (2013)
16. Simoens, P., Xiao, Y., Pillai, P., Chen, Z., Ha, K., Satyanarayanan, M.: Scalable crowd-sourcing of video from mobile devices. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, ser. MobiSys (2013)

17. Zhang, D., He, T., Raghu, F.Y., Zhang, G.D., Ganti, T.H.R., Lei, H.: Where is the crowd?: crowdedness detection scheme for mobile crowdsensing applications. In: IEEE International Conference on Computer Communications (INFOCOM) (2011)
18. Lane, N.D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., Campbell, A.T.: A survey of mobile phone sensing. Commun. Mag. **48**, 140–150 (2010)
19. Ra, M.R., Liu, B., La Porta, T.F., Govindan, R.: Medusa: a programming framework for crowd-sensing applications. In: Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services, ser. MobiSys (2012)
20. Das, T., Mohan, P., Padmanabhan, V.N., Ramjee, R., Sharma, A.: Prism: platform for remote sensing using smartphones. In: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, ser. MobiSys (2010)
21. Gao, H., Barbier, G., Goolsby, R.: Harnessing the crowdsourcing power of social media for disaster relief. IEEE Intell. Syst. **26**, 10–14 (2011)
22. Doan, A., Ramakrishnan, R., Halevy, A.Y.: Crowdsourcing systems on the world-wide web. ACM Commun. **54**, 86–96 (2011)
23. Yang, D., Xue, G., Fang, X., Tang, J.: Crowdsourcing to smartphones: incentive mechanism design for mobile phone sensing. In: Proceedings of the 18th Annual International Conference on Mobile Computing and Networking, ser. Mobicom (2012)
24. Han, K., Zhang, C., Luo, J.: Truthful scheduling mechanisms for powering mobile crowdsensing. CoRR (2013).arxiv:abs/1308.4501
25. Faridani, S., Hartmann, B., Ipeirotis, P.: What's the right price? pricing tasks for finishing on time. In: Conference on Artificial Intelligence (AAAI) (2011)
26. Hirth, M., Hossfeld, T., Tran-Gia, P.: Analyzing costs and accuracy of validation mechanisms for crowdsourcing platforms. Math. Comput. Model. (2012)

Vehicular Ad Hoc/Sensor Networks in Smart Cities

Chao Song and Jie Wu

Abstract Smart city comprises numerous technologies and depends on sensors to be aware of its environment. Vehicular sensing where vehicles on the road continuously gather, process, and share location-relevant sensor data (e.g., road condition, traffic flow) is emerging as a new network paradigm for sensor information sharing in urban environments. In this chapter, we introduce the vehicular network and the challenges in it. We will briefly discuss the existing routing protocols for vehicular networks, and analyze them by a macro–micro model. In addition, we will also cover the vehicular sensor applications in smart cities.

1 Introduction

One way to cope with challenges of modern urban environment (increasing population, pollution, energy consumption, etc.) is by making it intelligent [1]. Smart cities take advantage of the benefits of integrating citizens and cities into the natural and ecologically friendly environment using modern technologies to improve their lives. Urban governance can be achieved through appropriate responses to events, redistribution of resources based on new environmental conditions, or any unnatural problems such as accidental or catastrophic benefits of modern technology. The city highly integrates the meaning of the road and vehicular traffic infrastructure with the networks themselves. Thus, along with standard infrastructures such as water supply or electricity, smart urban development is highly dependent on the development of transportation infrastructures on roads and streets as they are all potential locations for sensor and data transmission paths. In addition, modern vehicles that

C. Song
School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China

J. Wu (✉)
Department of Computer and Information Sciences, Temple University, Philadelphia, USA
e-mail: jiewu@temple.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_9

are becoming smarter and more heavily equipped with sensors and actuators use road infrastructure. These already existing vehicle functions can be achieved by enabling vehicles to collect more general sensor data for further enhancement for data transfer, such as the heterogeneous mobile sensor network.

The combination of fixed and mobile sensors and networking technologies into agile, error-prone, modern, and powerful networks can prove valuable to the data infrastructure of the smart city. However, this network faces its unique challenges in which some nodes are stationary and some are mobile, and there is a basic requirement to determine the optimal routing path for data transmission. We wanted to achieve a constant presence of sensors in the cities and they did not have a fixed network infrastructure re-input interconnection. To achieve this goal, we should use a heterogeneous approach that will utilize all possible network access points to achieve dynamic interconnection of all possible intelligent devices and sensors.

Vehicle Ad Hoc Networks (VANETs) are in the process of acquiring a related businesses because of recent advances in inter-vehicle communication through the DSRC/WAVE standard [2] and stimulating brand new visionary service vehicles from entertainment applications to travel/advertising information, from driver safety to opportunistic intermittent connectivity and Internet access [3, 4]. In particular, Vehicular Sensor Networks (VSNs) are becoming a new tool to effectively monitor the physical world, especially in urban areas where high concentrations of vehicles equipped with onboard sensors are expected [5, 6], as shown in Fig. 1. In addition,

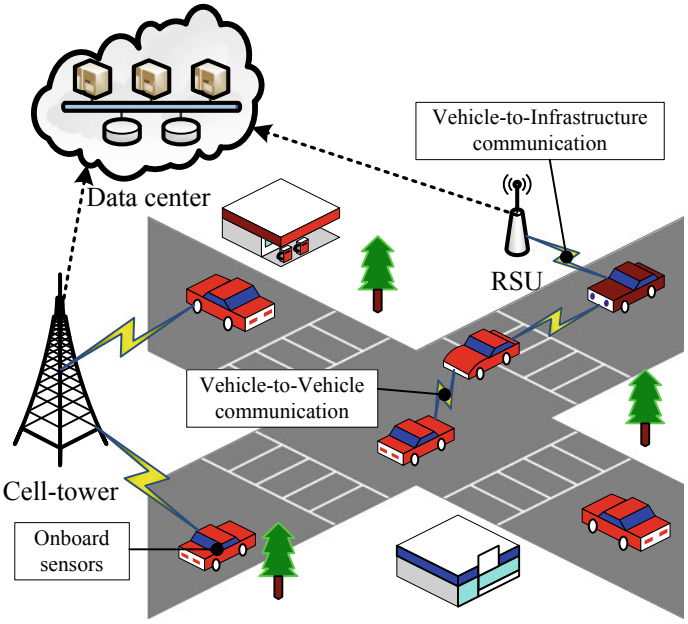


Fig. 1 The architecture of Vehicular Sensor Network (VSN)

with the onboard sensor and a 3G/4G mobile Internet connection, the growing popularity of smartphones sheds light on the use of smartphones as a platform involved in vehicle remote sensing [7]. As a result, the remote sensing platform for such vehicles will enable new applications such as street-level traffic estimation, ride quality monitoring, and active city surveillance.

Vehicles are usually not subject to stringent energy constraints and can easily be equipped with powerful processing units, wireless transmitters, and sensing devices even with some complexity, cost, and weight (GPS, cameras, vibration sensors, acoustic detectors, etc.). VSNs represent significant novelty and challenging deployment scenarios, and are significantly different from the more traditional wireless sensor network environments, and therefore, require innovative solutions. In fact, in a different way from conventional wireless sensor nodes, vehicles usually exhibit constrained movement patterns due to street layout, connections, and speed limitations. In addition, they usually have no strict limits on the processing power and storage capacity.

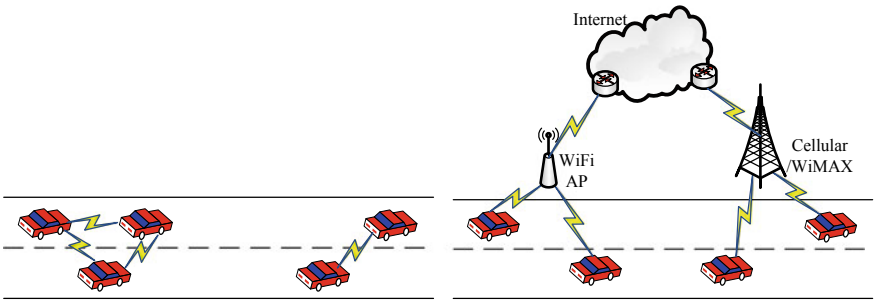
2 Background

2.1 Architecture

The Vehicle Sensor Network (VSN) platform provides a means for data collection/processing/access to the sensor. Vehicle sensor data is collected successively from city streets (for example, images, accelerometer data, and so on), and then processed to search for information of interest (for example, identification plate or to infer traffic patterns). Vehicle sensor information access network architecture depends largely on the wireless access method at the bottom of the vehicle environment. In general, the Vehicular Sensor Network (VSN) consists of three layers: a sensor layer, a communication layer, and a data process layer [8].

In the sensor layer, the vehicle is regarded as a large mobile sensor node. The vehicle is equipped with an Onboard Unit (OBU), which is an electronic device that can sense, communicate, and compute. It uses the OBU to sense the state of vehicle, such as the speed, moving direction, and location. OBU also senses the environment around the vehicle, such as the road traffic and climate. Moreover, most vehicles are equipped with a Global Navigation Satellite System (GNSS) device, which can offer positions and time synchronization for the vehicle, such as an American GPS or the Chinese BeiDou navigation satellite system (BDS).

The communication layer includes VANET, cellular network (3G or 4G), and mixed networks of the Internet. VANET uses the vehicles as mobile nodes in the MANET to create mobile networks. Every participating vehicle turns into a wireless router or node, allowing vehicles approximately 100–300 m around each other to connect and, in turn, create a network with a wide range. When the vehicle falls



(a) Vehicle-to-Vehicle (V2V) (b) Vehicle-to-Infrastructure (V2I)

Fig. 2 Communications in Vehicular Sensor Networks

in the signal range and when you exit the network, it can join other vehicles; the vehicles are then connected to each other, creating a mobile Internet.

If the vehicle is only equipped with vehicle communication equipment, it should operate in an infrastructure mode for free. Sensor data must be handled locally or in collaboration with vehicle-to-vehicle (V2V) communications [9, 10] to facilitate access to information, as shown in Fig. 2a. Roadside Units (RSU) are the fixed infrastructures which can help in forwarding the data packet to promote reliable communications. If the vehicle is also equipped with access methods, such as 3G/4G and WiMAX broadband wireless access, it can use a vehicle-to-infrastructure (V2I) communication sensor data sharing over the Internet. Mobile users can report the sensor's data to the Internet servers, and other users can access information from these servers [6], as shown in Fig. 2b.

The data processing layer includes two parts,namely, the data storage part and the data analysis part [8]. A lot of the vehicles' information, traffic data, and other data require the appropriate storage mechanism. Google's product is a stable solution, and is divided into three levels from the bottom up—basic file system (Google file system or the Colossus), database management (BigTable), and the programming model (MapReduce). Data analysis is important for the applications of VSN, which provides functions such as location-based services and vehicle monitoring services.

2.2 Environment

The impact of the environment on VANET is significant. The strategy of vehicular network under different environment will be different. The road structure under urban environment is regular, such as the grid road structure in Manhattan. The population density in urban environment is very high. Moreover, the vehicles in urban environment are driven at low speed due to the speed limit. Therefore, the data transmissions of the vehicular network under the urban environment are often in the inter-road mode, that is, the data packets are delivered along the multiple roads. On

the contrary, the roads under rural environment are not structured, including several intercity roads and highways. Moreover, the traffic volume is relatively low. Because of the sparseness of roads and vehicles, the data transmissions of the vehicular network under the rural environment are often in the intra-road mode, where data packets are usually delivered among vehicles in the road, such as highway.

2.3 DSRC/WAVE Protocol Stacks

Dedicated short-range communication (DSRC) is working in a 5.9 GHz medium-distance communication technology of a short distance [2, 11]. Standards Committee E17.51 recognized a variant of the IEEE 802.11A MAC DSRC link. DSRC supports vehicles at speeds of up to 120 mph and a nominal range of 300 m (up to 1000 m), the default is a 6 Mbps data rate (up to 27 Mbps). This will be referred to as DSRC/WAVE (Wireless Access in a Vehicular Environment) in a variety of application environments to achieve and improve traffic flow, highway safety, and other Intelligent Transportation Systems' (ITS) application-related operations. DSRC has two modes of operation: (1) Ad hoc mode, which is characterized by distributed multi-hop networking (vehicle–vehicle), and (2) Infrastructure mode,which is characterized by a centralized mobile single-hop network (vehicle-gateway). Note that, depending on your deployment scenario, the gateways can be connected to each other or connected to the Internet, and they can be equipped with computing and storage devices, such as the Infostations [12].

3 Unique Challenges of Vehicular Networks

In vehicular ad hoc/sensor networks, the vehicles with radio ranges are regarded as the mobile nodes and routers to other nodes. The vehicular networks are similar to the traditional ad hoc networks, such as self-management, short radio transmission range, self-organization, and low bandwidth. However, the vehicular networks have unique challenges that affect the design of the routing protocols [13]. These challenges include the following:

High-speed mobility: The nodes in the traditional ad hoc networks (such as wireless sensor networks) are often stationary, and of course there are some sensor nodes that are moving, but the speed is slow,with the speed range of 1–5 m/s. In the vehicle sensor network, the node is in the city or highway moving vehicles, the speed range is usually 10–30 m/s, or even higher, so the node has high-speed mobility. Most existing researches in sensor networks assume that the entire network is connected after the deployment of sensor nodes, and any network node can be found in the network topology to a data sink node path. However, in the vehicle sensor network environment, this assumption is not established.

Intermittent connectivity: Intermittent connectivity in the network refers to the internode communication in the network that cannot guarantee a stable and continuous connection for a period of time, the connection is always intermittent. In [14], it is discussed that if the coverage radius of the node is 250 m at an average speed of 100 km/h, the probability of the link being 15 s is only 57%. The high-speed mobility of nodes can also cause rapid changes in network topology. In the car network, due to the high-speed mobility of nodes, the topology changes frequently. Network connectivity has a large impact on communication protocols, but the in-vehicle network is an intermittent connectivity network, making it difficult to establish end-to-end reliable connections.

Frequent changes in the topology: In the traditional ad hoc networks (such as wireless sensor networks) research, the vast majority is based on the traditional self-organizing network characteristics, assuming that the nodes are connected in the network, that is, each node in the network topology can find an end-to-end routing path to another node for data transmission. However, it is difficult to find such stable end-to-end routing path in a vehicular network. In the vehicular network, it is difficult to solve the problem of routing in the network by establishing a relatively stable routing table like the traditional network to manage the topology between nodes.

Opportunistic data delivery: In VANETs, the connectivity between nodes change frequently with the mobility of nodes, and the event of forwarding data occurs when a node meets with other nodes (in the range of communication radius of each other). Thus, the data transmission adopts the way of store and forward, that is, when the two nodes meet, they establish a wireless connection for exchanging information and storing data, and then forward this data to the next node encountered until the transfer to destination. However, the vehicle in the network does not understand the future travel route of its neighbors, so the uncertainty of such vehicle movements will lead to the way in which the mobile node-based data is forwarded. One solution is to copy the data and forward it to more vehicles in a multipath way to the target point, although the more data to be replicated to the neighbor vehicle will improve the efficiency of data transfer, but the vehicle sensor network resources are limited. This resource includes the network bandwidth and the network of mobile nodes in the cache space, if each car has copied a lot of data, then the network of these limited resources will soon be exhausted, but will affect the performance of data transmission.

In addition, we take the traffic hole problem as an example to illustrate the unique challenges VANETs [15]. The problem of traffic holes can be seen everywhere in the traffic environment. Even during peak hours, the traffic volume is the highest. The road shown in Fig. 3 is the Chengdu Bust Road, which is the busiest in Chengdu, and is the city with more than 2 million cars in China. The photos were taken in the afternoon peak hours. Initially, the traffic flow on this road was saturated, as shown in Fig. 3a. However, after 2 min, as shown in Fig. 3b, the traffic flow dropped sharply. After 3 min, as shown in Fig. 3c, there is no vehicle on this road and there is a gap. All vehicles on the road were blocked by traffic lights at the entrance. If the length of this gap is greater than the communication range (R) of the vehicle, the gap will block wireless communication between the vehicles.

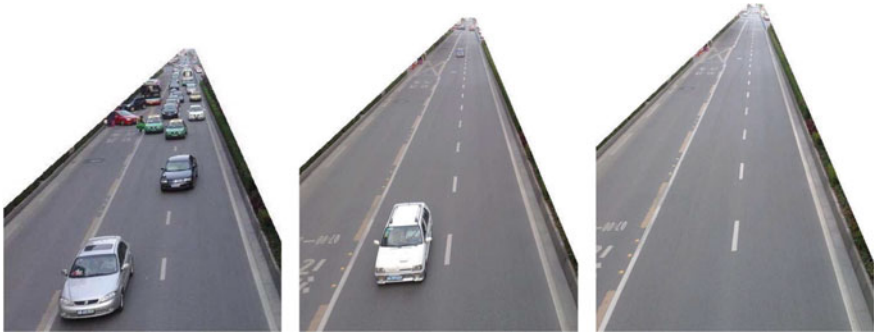


Fig. 3 Appearance of a traffic hole

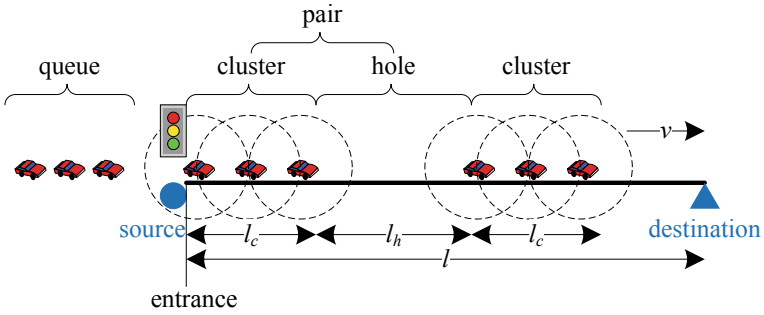


Fig. 4 The traffic hole and connected cluster on a road

The traffic flow which is regulated by the interaction among the vehicles and the roadways is called *uninterrupted flow*, such as a vehicle traveling on an interstate highway. In contrast, traffic flow which is caused by external means such as traffic lights or pedestrian signals is the term used as an *interrupted flow*. Figure 4 shows an example of the disrupted traffic flow along a road. The source of the sent data packet is at the road entrance, which is also a signal intersection. The destination is at the exit of the road or road. The distance between the source and the destination is l.

Let *traffic hole* be the gap in a traffic flow on the road. As shown in Fig. 4, its length (denoted by l_h) is larger than the communication range of the vehicles, i.e,. $l_h > R$. Let the *connected cluster* be a group of vehicles on the road that can communicate with each other via either one-hop or multi-hops communication, and its length is denoted by l_c, as shown in Fig. 4.

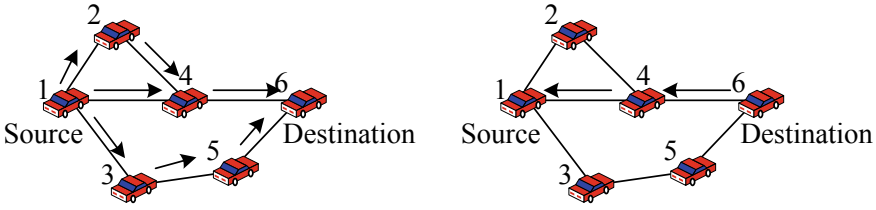
4 Routing in VANETs

In view of the different architectures, applications, and challenges, researchers have made a wide range of routing protocols for VANETs [13]. These protocols are mainly aimed at maximizing throughput while minimizing packet loss and control costs. One of the main challenges in VANET is the development of an efficient routing protocol for a highly variable topology. VANET needs new types of routing protocols. Opposite of the wired infrastructure, it does not use a dedicated router node. The routing protocols are used by the user node (vehicle), which can be mobile and unreliable. The current routing protocol for VANET can be divided into two main categories: vehicle-to-vehicle-based (V2V) routing protocols and vehicle-to-infrastructure-based (V2I) routing protocols. V2V protocols perform vehicle-to-vehicle communication but do not focus on fixed infrastructures on roads. It can be divided into four groups: (1) topology-based (ad hoc) routing protocols, (2) position-based routing protocols, (3) cluster-based routing protocols, and (4) broadcast-based (geocast) routing protocols. Moreover, the deployment of a communications infrastructure and the road to vehicle communication is more reliable and reduces unwanted delays in the application of different vehicles.

4.1 Topology-Based (Ad Hoc) Routing Protocols

Topology-based routing uses existing communication links to forward packets. Ad hoc On-demand Distance Vector (AODV) [16] is a refinement of the DSDV Protocol. AODV differs from DSDV because it reduces the number of broadcast messages and the routing is created on-demand. However, DSDV maintains all the routes listed in the routing table.

The source node in AODV uses a Hello Beacon to detect its neighbor to start the routing protocol. As shown in Fig. 5a, to find the path to the destination, the source node broadcasts a Route Request Packet (RREQ), and then its neighbor broadcasts the RREQ to its neighbors until it reaches the route to the destination node. After



(a) Propagation of RREQ (b) RREP's path to the source

Fig. 5 AODV routing protocol

receiving the RREQ packet, the node will register the address of the node sending the query in its routing table. The method of registering the previous hop is called learning later. As shown in Fig. 5b, after reaching the target node, the Route Reply Packet (RREP) is transmitted through the path from the back to the source.

When the source node moves, the route to the destination is established. If one of the intermediate nodes moves, then its neighbor realizes the link failure and sends a notification of the link failure to its upstream neighbors, and so on until it reaches the source. Thus, if needed, the source can reinitiate the process of route discovery again. Due to the periodic beaconing, the protocol tends to run out of the extra bandwidth. In addition, a single RREQ with multiple RREP packets will cause a serious control overhead.

4.2 Position-Based Routing Protocols

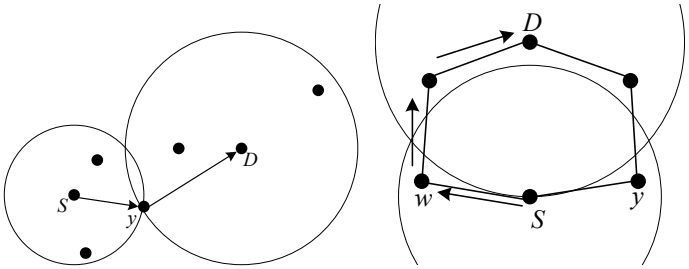
In the position-based routing protocol, all nodes point devices, such as GPS location, to identify their own position and their neighbors geographically [16]. It does not manage any routing table or exchange associated with the neighbor link state information. Information from the GPS equipment is used to make routing decisions. This type of routing performs better, because creating and maintaining global routing from the source node to the destination node is not necessary. Location-based routing protocols can be classified as non-delay-tolerant network (non-DTN) routing protocols, and delay-tolerant network (DTN) routing protocols.

4.2.1 Non-delay-Tolerant Networks (non-DTNs) Routing Protocols

Non-DTN routing protocols do not use the alternating connectivity, and are valid only in sufficiently populated vehicular networks. These protocols are designed to deliver data packets to the destination as soon as possible. The basic idea of the greedy method of non-DTN routing protocols is that a node sends its data packet to one of its neighbors that is closer to the destination. However, if the neighbors are not closer to the destination than that node, the forwarding policies may not be successful. Therefore, we can claim that the routing protocol has reached the local maximum at the node, because it has achieved the maximum local growth at the current node. The routing protocols in that group have their own recovery method to handle such a failure.

Greedy Perimeter Stateless Routing (GPSR) [17] is a location-based routing protocol aimed at addressing the mobile environment. GPSR is most suitable on the highway in which nodes are uniformly distributed. This routing protocol depends on the following two modes:

Greedy mode where the node forwards a data packet to one of its neighbors that is closer to the destination node by considering the position of the neighbors in the network topology. As shown in Fig. 6a, the node *S* wants to send a data packet



(a) Greedy forwarding mode (b) Perimeter forwarding mode

Fig. 6 GPSR routing protocol

destined for node D. S, sends the data packet to the node y, which is in S's neighbors and is closer to D than any of S's other neighbors.

Perimeter mode is used as the protocol recovery mode when the packet reaches a local maximum. Therefore, the recovery mode is used to forward packets to nodes that are closer to local destinations than where the packets are facing local maximums. A simple example of this topology is shown in Fig. 6b. The node S is closer to the destination node D than its neighbors w and y. S does not choose to forward data packets to w or y using greedy forwarding, although the two paths to D are $S - y - z - D$ and $S - w - v - D$. In this case, the GPSR protocol declares S as the local maximum to D, and the shaded area has no nodes, as in the invalid area. To route packets around invalid areas, the perimeter of the forwarding strategy constructs a planarized graph for the neighbor node S and routes the invalid packets around the packet with the right-hand rule. The right-hand rule states that when the node reaches from y to S, the traversed edge is a counterclockwise rotation around S from edge (x, y). By applying the right-hand rule shown in Fig. 6b, the data from node S forwards the packet to the hop w.

4.2.2 Routing Protocols for Delay-Tolerant Networks (DTNs)

DTN is a method of solving computer network architecture problems in a heterogeneous network that may lack continuous network connectivity, and consequently, lacks instantaneous end-to-end paths. Examples of such networks are those that operate in mobile or extreme terrestrial environments or planned cyberspace. The development of the routing protocol of the vehicle has been regarded as a kind of DTN characteristic. Given the challenging environment of such networks, they are subject to a periodic connection loss. In order to solve this problem, the packet transmission is increased by allowing nodes to store data packets when they lose contact with other nodes, and put the data packets at a certain distance as long as it satisfies the other nodes according to some indicators with the neighboring nodes; this is called the carry-and-forward strategy.

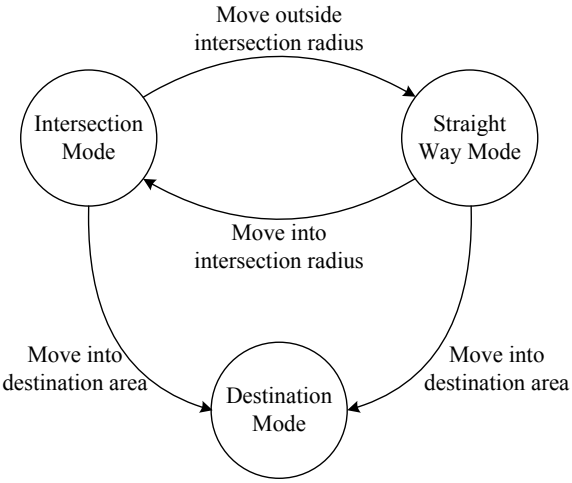


Fig. 7 The transition modes in VADD

Vehicle-assisted data delivery (VADD) [18] is a routing strategy for vehicles that is intended to enhance the routing of vehicles based on the concept of carry-and-forward by utilizing the mobility of vehicle. A car makes a decision at an intersection while selecting the next forwarding path with a negligible delivery delay. The path is an intersection where only the split path is split. The best path for packet forwarding is switched between the three packet modes (Intersection, Straight Way, and Destination). As shown in Fig. 7, VADD has three packet patterns: Intersection, Straight Way, and Destination based on the location of the packet carrier (i.e., the vehicle which carries the data packet). By switching between these packet modes, the packet carrier needs the best packet forwarding path. Between the three modes, the Intersection mode is the most critical and complex one, because the vehicle at the intersection has the most choices.

4.3 Cluster-Based Routing Protocols

In general, cluster-based routing protocols are more suitable for network cluster topology. Each cluster has a cluster head for intercluster management purposes. The intra-cluster nodes interact with each other through direct contact, while cluster-to-cluster interactions are performed through the cluster headers. In a cluster-based routing protocol, the clusters are formed close to each other. However, in cluster-based routing protocols, cluster configuration and cluster head selection is an important issue. Due to the high mobility of VANET, dynamic cluster configuration becomes a major process.

In the Cluster-Based Directional Routing Protocol (CBDRP) [19], vehicles traveling on the same route are split into several clusters. Each vehicle can communicate

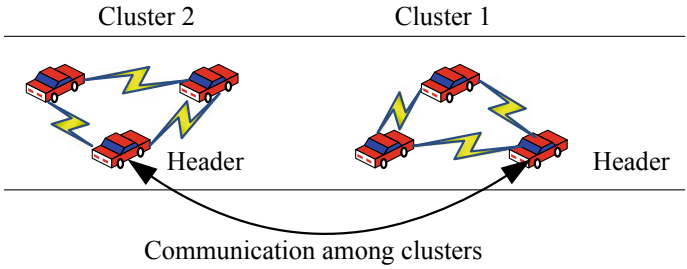


Fig. 8 Cluster splitting example in CBDRP

with its neighbors in the same cluster by radio. An example of cluster splitting is shown in the Fig. 8. The figure shows two clusters. The center of a cluster is fixed after it partitions. The CBDRP protocol assumes that the radius of the radio is r, the length of each cluster is d, and the half-width of the road is w. Given that $d > w$, d is almost equal to $r/2$. The radio transmission radius of 802.11 p is 1000 m, so the theoretical length of a cluster can be as high as 500 m. D can be much larger if the header is near the center. The source node in CBDRP forwards the message to its cluster head, and then sends the message to the header, which is located in the same cluster as the destination. Eventually, the destination header forwards the message to the destination. Select the cluster header; the persistence is similar to CBR, but it considers the speed and direction of the vehicle. Simulation results show that CBDRP can solve the link stability of the vehicle, so as to ensure reliable and fast data transmissions.

4.4 Broadcast-Based Routing Protocols

Broadcast-based routing is commonly used by VANET to share information about traffic, weather, and emergencies, car usage, and advertising and announcements for the [20]. Broadcast-based routing protocols follow simple broadcast flooding in which each node resends to other nodes. This process guarantees the arrival of all destination messages, but with a high overhead. In addition, it is only suitable for a large number of smaller nodes in the network. A larger node density results in more message broadcasts resulting in collisions, higher bandwidth utilization, and overall system performance degradation.

Urban multi-hop broadcast (UMB) [21] aims to address (i) broadcast storms, (ii) hidden nodes and (iii) reliability problems in multi-hop broadcasts in urban areas. This protocol assigns forwarding and acknowledgment to only one vehicle. It splits the sections of the road inside the transmission range into portions, and selects the vehicle in the furthest non-empty segment without a priori of topological information. When there is an intersection in the path of the message propagation, a new directed broadcast is initiated by the repeater at the intersection.

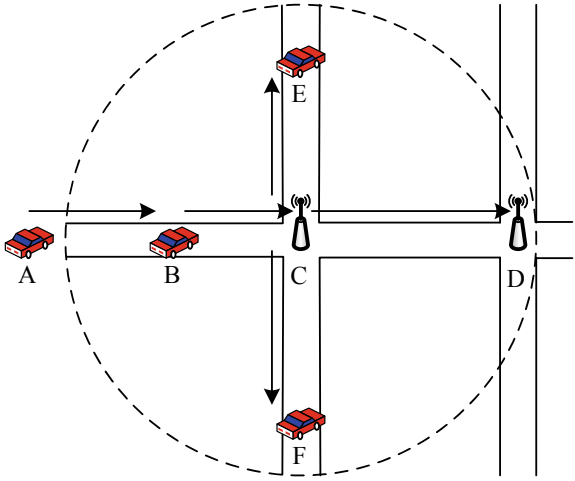


Fig. 9 Intersection handling in UMB protocol

Figure 9 illustrates an example of intersection handling in UMB, where the directed broadcast is used in the case where the vehicle A reaches the vehicle B. Vehicle A is out of the transmission range of repeater C. At the same time, the vehicle B is in the transmission range of the repeater C. Therefore, the vehicle B uses the IEEE 802.11 protocol for the communication repeater C. Once the repeater receives the message, it initiates a directed broadcast of the north and south directions. Since the repeater D is in the transmission range of C, C also sends the packet to the repeater D by using the IEEE 802.11 protocol.

In addition, data delivery in VANETs can be divided into two categories, road delivery and delivery delivery. Since traffic is two-way, data delivery on roads is much simpler than delivery. Therefore, the opportunistic forwarding through intermittent connections between intersections and vehicle nodes is a challenging issue in VANETs. In order to overcome this problem, a Buffer and Switch protocol (BAS) [22] was proposed. The basic idea behind this is to use the vehicle nodes on each road to buffer multiple propagated packets in order to provide more opportunities for packet switching at the intersection. Due to resource limitations in VANETs (such as vehicle node bandwidth and storage space), BAS employs space–time controlled repeat propagation on the routing path. Unlike conventional protocols in VANETs, duplicates in the BAS propagate not only to the preceding vehicle node (called *downstream propagation*) but also to the nodes along the routing path (called *upstream propagation*). For the effective packet switching at each intersection, each packet on the previous road must be: (1) Using greedy forwarding protocol to timely forward to the intersection; (2) Spread to more vehicle nodes to provide more opportunities for effective packet switching at the crossroads. We call it the way to buffer packets.

Therefore, the propagation on the road consists of two phases: (1) Downstream propagation: In order to send packets in time to the expected next intersection, each packet transmitted on the node is copied to its neighboring location and is closest to

the destination along the routing path, and any of its neighbors hold this package in front of it; (2) Upstream Propagation: For buffering packets on the road, each packet carried in the node is also copied to its next adjacent location, which is the furthest location along the routing path and behind it. In addition, in order to reduce resource consumption, this spread is limited on this road.

4.5 Infrastructure-Based Routing Protocols

The placement of a fixed RSU, linked to the exact location of the backbone, is a necessary condition for communication. The number and distribution of RSUs depends on which communication protocols have to be adopted. For example, some protocols require that RSUs be evenly distributed throughout the entire road network, while some others only need to be at intersections, and others only need to be at the regional boundaries. It can be assumed that the infrastructure is beaten to a certain level and the vehicle occasionally has access to it.

RSUs in VANET offer two potential benefits: In the first case, the higher antenna height increases the reliability of the Vehicle-to-Infrastructure (V2I) communications compared to Vehicle-to-Vehicle (V2V) communications. In addition, the deployment of RSUs are connected to a higher bandwidth and a more reliable backbone network, providing traffic management departments with centralized access, enabling configuration, and maintenance of these units.

The routing protocol of Static node-assisted adaptive (SADV) [23] is designed to minimize the message delivery delay in sparse networks, and adjusts to varying traffic densities by enabling each node to estimate the amount of the message delivery delay. SADV assumes each single vehicle and knows its location via GPS, and each has access to an external static street map. There are three different modules for SADV: (1) Static Node-Assisted Routing (SNAR), (2) Link Delay Update (LDU), and (3) Multipath Data Dissemination (MPDD).

SADV operates a road mode and an intersection mode. SNAR determines the optimal path by using the graph based on the abstract from the roadmap. The LDU maintains a delay matrix that dynamically measures the delay between the message-passing static nodes. MPDD assists in multipath routing. Static nodes in the SADV have a similar concept, such as Throwbox, which can store and forward data in the DTN routing protocol, where the SADV can store and forward the data necessary for the device. However, SADV is different from regressive because SADV does not require the node contact option, but requires the implementation of static nodes and routing algorithms to take advantage of the street map structure and vehicle density for each road in the vehicle network.

The static node may store the packet for a period of time until the shortest delay path becomes available. As shown in Fig. 10, the packet is sent from A to the remote location. Once the packet is transferred from A to B, the latter has determined that the packet is to be forwarded to the next vehicle. For example, it is assumed that the shortest delay path to the packet is to the north, but no traffic within the

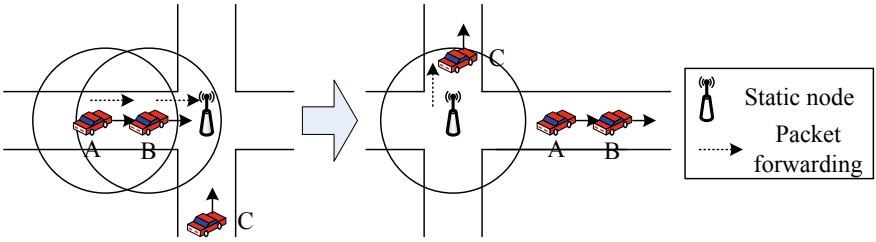
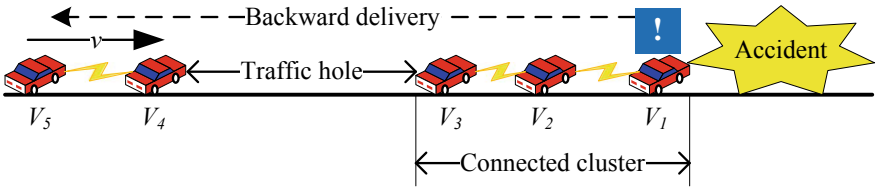
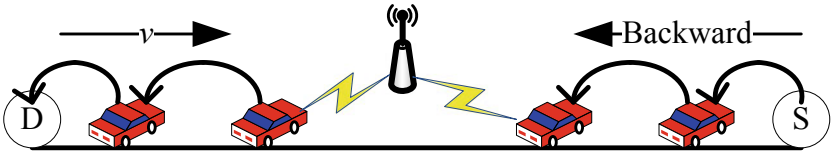


Fig. 10 SNAR mode in SADV routing protocol



(a) The traffic hole problem



(b) RSU-Assisted Backward Delivery

Fig. 11 Backward data delivery in VANETs

communication range of the vehicle *B* can transmit the packets along this direction. Therefore, *B* transmits the data packets to the static node. The static node stores the packet for a period of time and sends it to *C* as it crosses the intersection to north of the road, as shown to the right of Fig. 10. Without support from the static node, the packet will be moved via *B* to the east of the road. However, if it does not meet *C* at the intersection, it may take a fairly long time for the packet to travel.

Another example is the backward data delivery in VANETs. Many applications in VANETs require delivering data packets backwards. In sensing applications, vehicles need to obtain the conditions about the roads that lie ahead of them. As an example in Fig. 11a, an alert message needs to be delivered from the first vehicle, V1, to the 5th vehicle, V5. The total data delivery delay is calculated from the interval time between when the first vehicle generates the alert message and when V5 receives the message.

Data delivery with V2V communications in VANETs is based on the vehicles on the roads, but the distribution of the vehicles could be affected by their mobilities

or by external means, such as traffic lights. A gap with a distance larger than the communication range of the vehicles could appear along the traffic flow; this is considered a traffic hole. It could block the data delivery along the traffic flow. As shown in Fig. 11a, when the distance between the vehicles V3 and V4 is larger than their communication range R, a traffic hole appears in the road traffic flow and partitions the road traffic flow into two connected clusters. On a one-way road, the message is backwardly delivered from the vehicle V3 to V4, and is blocked by the traffic hole. Because the data are headed in the direction opposite to that of the motion of the vehicle, no available vehicles can carry the data to the destination using the movement-assisted routing protocol.

We can utilize some static roadside units (RSUs) to help forwarding the packets to reduce the data delivery delay, as shown in Fig. 11b. An RSU can be a wireless access point, a parked vehicle, vehicles waiting at an intersection, or the static node. The RSU only acts as a relay, so it incurs a lower cost than a traditional access point. We term this type of data delivery as RSU-Assisted Backward Delivery (RABD) [24]. Each road can be regarded as a river. The vehicles are regarded as boats, and they can move from upstream to downstream. Thus, each RSU can be seen as a dock for delivering the data packets.

5 Macro–Micro Model for Routing Protocols

Without considering the details of highly dynamic network topology, we propose a model called Macro–Micro model based on our understanding of the VANET characteristics to analyze and resolve the routing problems in VANETs. Macro–Micro model divides the vehicular network into macroscopic level (Macro) and microscopic level (Micro). The Macro provides information for computing routing strategies in VANETs, and the Micro defines the data delivery protocols which guarantee packets are effectively delivered to their destinations.

5.1 State Awareness Routing Protocols

Depending on state awareness, routing protocols in VANETs can be classified into three categories, namely, state-aware, stateless, and hybrid routing protocols. State-aware protocols, such as DSR [25] or AODV [16], maintain a routing state (routing table or routing path) in the communicating nodes. The state-aware protocols are not adapted to VANETs due to the short duration of the routing states and the overhead of their maintenance. Due to high mobility, the topology of VANETs changes rapidly. Such particular features often make these protocols inefficient or unusable in VANETs. Therefore, a direct approach is to develop stateless routing protocols, i.e., protocols that do not maintain routing states, such as greedy routing GPSR [17]. However, the researchers in [18, 26, 27] have argued that the greedy routing lacks

knowledge of topology and may fall into a local minimum. Although a recovery mode can be used to escape from the local minimum, it is not efficient if packets spend more time in the recovery phase than in the greedy phase. Hybrid protocols use a combination of state-aware and stateless approaches. Since vehicles should try to make use of the wireless communication channel as much as possible, packets will be transmitted along the path with higher road traffic density. Previous researches [18, 27] use density and road lengths as the metric for route creation which is state-aware, and deliver packets by some greedy protocols which are stateless.

However, they haven't considered the network congestion which could seriously affect the packet delivery delay of each road. Because all vehicles use traffic density as metric for route creation, most packets will be delivered along those roads with higher density. With the increase in the number of packets, because of limited resources (buffer and bandwidth), network congestion may arise in these roads, and thus decreasing the protocols performance in terms of delivering packets. However, traffic densities of these roads can not be affected by the network congestion, so vehicles will continue to deliver packets to them. This can reduce the performance of packets delivery in the network.

5.2 Macro–Micro Model

In [28], the authors propose that due to the lack of knowledge of network topology evolution, the design of effective routing strategies for opportunistic networks is usually a complex task. When more knowledge about the expected topology of the network is available, routing performance will increase [28, 29]. Unfortunately, this knowledge is not readily available, and tradeoffs must be made between performance and knowledge needs. Therefore, in Macro–Micro model, we have studied the collection of more valuable routing information in macros. This macro has state awareness, which relates to the services provided by the road, such as packet delivery delay or travel time, and is typically used to calculate packet routing strategies or vehicle travel plans. The routing policy may be the same as a single routing path (such as GPSR [17]) or multiple routing paths (such as epidemic routing [30]).

In Micro level, we investigate the details of data packets delivery in the vehicular network with two necessary units, which are the vehicle node (or Roadside Unit, RSU) and the data packet. The Micro level analyzes and manages the interactions between the vehicle nodes and the data packets. The moving vehicles in VANETs are equipped with wireless onboard units (OBUs), which communicate with each other or RSUs by a dedicated short range communications (DSRC) [4] protocol. For analyzing the vehicle nodes, we consider two properties of vehicles, which are mobility and communication. The movement of vehicles is directional, and the mobility property of a vehicle node includes speed, acceleration, direction, GPS position and so on. Some studies [18, 26] have presented that vehicle nodes in a vehicular network can periodically broadcast HELLO beacon message about their mobility properties to

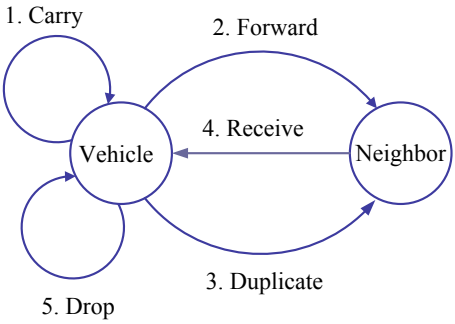


Fig. 12 State diagram of data packet

their neighbors, and each node can obtain the mobility information of its neighbors by the beacon messages.

The communication property of a vehicle node includes the protocols in the physical layer, the data link layer, the network layer and so on. A vehicle node can communicate with other nodes by some wireless communication devices, and the vehicle nodes interact with packets through this property. In Fig. 12, we present the possible states of data packets and the associated transition diagram, and we classify the interactions between the vehicle nodes and the data packets into five categories: (1) Carry: the vehicle node stores the packet in its local buffer and moves; (2) Forward: the vehicle node forwards the packet to another node through wireless communication; (3) Duplicate: the node duplicates the packet and forwards a copy of it to another node; (4) Receive: the node receives the packet from one of its neighbors; (5) Drop: the node deletes the data packet from its buffer. Based on these categories, various protocols can be used for data delivery in Micro, such as greedy forwarding [17], epidemic routing [30] and so on.

5.3 Mapping in Macro–Micro Model

We can map the categories of state awareness routing protocols to Macro–Micro model, and analyze them based on it. As shown in Fig. 13, the x-axis denotes the awareness of states, and the y-axis denotes the levels in VANETs which are the Macro and the Micro. State-aware protocols (such as AODV and DSR), which are regardless of state in Macro and are state-aware in Micro, are mapped to the quadrant II and IV. Stateless protocols (such as GPSR), which are regardless of state both in Macro and Micro, are mapped to the quadrant II and III. Hybrid protocols (such as VADD and LOUVRE), which are state-aware in Macro and stateless in Micro, are mapped to the quadrant I and III. These protocols are summarized in Table 1.

As we have discussed, the state-aware and stateless protocols cannot effectively solve routing problems in VANETs. However, some hybrid protocols [18, 27] deliver packets with greedy forwarding protocols which are based on some stable statistics

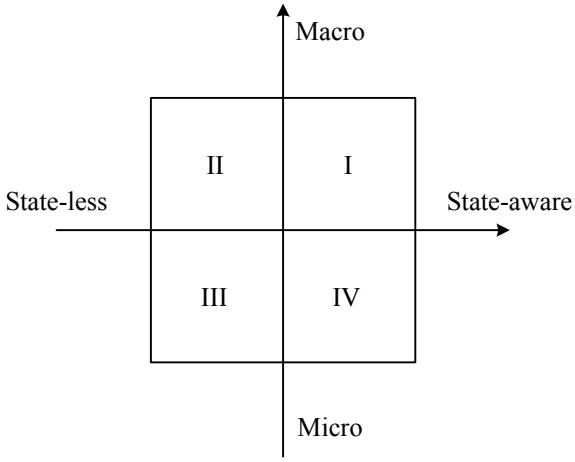


Fig. 13 Mapping in Macro–Micro model

Table 1 Overview of the routing protocols in Macro–Micro model

Level	Micro		Macro	
State	State-aware	Stateless	State-aware	Stateless
DSR [25]	√			√
AODV [16]	√			√
GPSR [17]		√		√
Epidemic [30]		√		√
VADD [18]		√	√	
LOUVRE [27]		√	√	
BAS [22]		√	√	

such as traffic density. We will utilize Macro–Micro model to analyze the effectiveness of these protocols.

VADD in [18] utilizes the estimation of packet forwarding delay through each road, which is based on some statistical data such as the average vehicle density, for route creation. Then, it proposes a greedy protocol to deliver packets. From Macro–Micro models perspective, VADD also includes two levels: the Macro is the estimation of packet forwarding delay of each road, and the Micro is the greedy forwarding protocol. As we have introduced, the purpose of Macro is to provide effective information for route creation. However, VADD is based on preloaded statistics which cannot adapt to the changing conditions in VANETs. Moreover, because it utilizes average vehicle density to estimate packet forwarding delay of each road, this may result in inaccurate estimation and cause the network congestion. As a result, the performance of VADD may be reduced.

Like VADD, LOUVRE [27] also uses density and road lengths as the metric for route creation in Macro. And the authors assume that vehicles are uniformly

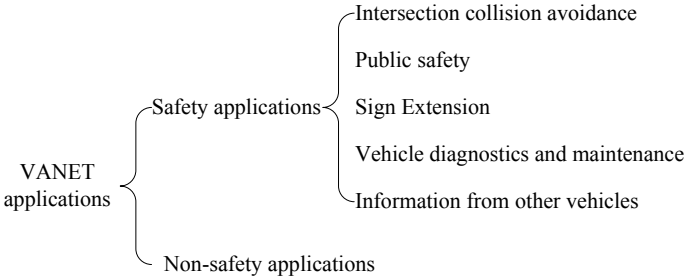


Fig. 14 Categories of VANET applications

distributed along a road. As long as the number of cars on the road is greater than or equal to a density threshold, the authors consider that the road is connected for routing. However, this assumption may not hold in a real VANET with highly dynamic topology and nonuniform node distribution.

6 Vehicular Sensing Applications

The communications of vehicle-to-vehicle and the vehicle-to-infrastructure in VANETs can support a large number of applications which require the data transmissions among the users or the devices [31–34]. Different types of sensors and GPS receivers are integrated with the network interface onboard devices that grant the vehicle the ability to collect, process, and disseminate information about itself and other vehicles approaching it in its environment. This has led to improved road safety and the provision of passenger comfort to [35, 36]. As shown in Fig. 14, the VANET applications are classified as safety applications and non-safety applications [37].

6.1 Safety Application

The safety applications use wireless communication between vehicles or between vehicles and infrastructures to improve road safety and avoid accidents; its intention is to save human lives and provide a clean environment. Safety applications are used for their ability to collect information from vehicle sensors, from other vehicles, or both, in order to process and disseminate information to other vehicles or infrastructures depending on the application and its function. The application of wireless communication technology in vehicles to communicate with other vehicles or infrastructures enables a wide range of applications and leads to increased road safety levels.

Safety applications that use V2V and/or V2I communications can be classified as [37]: (1) intersection collision avoidance, (2) public safety, (3) sign extension, (4) vehicle diagnostics and maintenance, and (5) information from other vehicles. For example, [38] propose the application of warning about violating traffic signals. This application is designed to send a warning message to the vehicle and warn the driver that a dangerous situation (accident) will occur if the vehicle does not stop; when the traffic light is running, signaling to stop sending messages depends on several factors such as traffic status, time, speed, vehicle location, and road surface. Improving the intersection collision avoidance system will lead to avoiding many road accidents, and the system is based on V2V or V2I communications.

6.2 Non-safety Application

The non-safety applications are designed to enhance the comfort of drivers and passengers (making the journey more enjoyable) and improve traffic efficiency [37]. They can provide drivers or passengers with weather and traffic information and detailed information on the location of the nearest restaurants, gas stations, or hotels, and their prices. Passengers can play online games, access the Internet, and send or receive instant messages, although the vehicle is connected to the infrastructure network.

The comfort of rides has been identified as one of the highest standards affecting customer satisfaction with public transport systems; thus, comfort is an important consideration for passengers using public transport. In particular, some passengers (such as pregnant women, children, and patients) need a more comfortable riding experience while traveling. The factors affecting passengers' ride comfort include: (1) individual factors such as driver behavior or vehicle condition; and (2) road conditions that affect most vehicles on a road.

Based on smartphones, a system [39] named Riding Experience Sensor (RESen) is proposed to sense and evaluate the riding experience. With the help of participatory phone sensing, RESen harvested a riding experience while driving and classified the experience horizontally and vertically. In order to adapt to a variety of different phone configurations, RESen can feel the horizontal and vertical experience and any direction. Based on the collection of participatory sensor data, RESen can evaluate the riding experience of three levels (track, road and driver). For trajectories, RESen should provide an overall riding experience, including anomalies along it. For a path, RESen will evaluate its riding experience, based on the track that passes it. RESen can evaluate the driver by comparing his trajectory and the riding experience of the road. Based on the evaluation, RESen can not only improve the driver's behavior, but also provide the user with a comfortable travel plan.

7 Conclusion and Future Research Directions

In this chapter, we studied the Vehicular Ad Hoc/Sensor Networks in Smart Cities. We presented the architecture and DSRC/WAVE Protocol stacks in vehicular sensor networks. We discussed the unique VANET challenges. We introduced the different kinds of routing protocols in VANETs. Moreover, we presented the vehicular sensing applications. In the future, we believe that new routing protocols can be provided for vehicular sensor networks, as well as other factors involved in our discussion for different kinds of applications in smart cities. In general, the field of vehicular network (VANET or VSN) is a new and growing area, which expects good perspective for the future.

References

1. Tomaš, B., Vrček, N.: Smart City Vehicular Mobile Sensor Network, pp. 70–77. Springer International Publishing, Cham (2015)
2. Lee, U., Gerla, M.: A survey of urban vehicular sensing platforms. Comput. Netw. Int. J. Comput. Telecommun. Netw. **54**(4), 527–544 (2010)
3. Nandan, A., Das, S., Pau, G., Gerla, M., Sanadidi, M.Y.: Co-operative downloading in vehicular ad-hoc wireless networks. In: Conference on Wireless On-Demand Network Systems and Services, pp. 32–41 (2005)
4. Xu, Q., Mak, T., Ko, J., Sengupta, R.: Vehicle-to-vehicle safety messaging in DSRC. In: Proceeding of International Workshop on Vehicular Ad Hoc Networks, pp. 19–28 (2004)
5. Lee, U., Lee, J., Park, J.S., Gerla, M.: FleaNet: a virtual market place on vehicular networks. IEEE Trans. Veh. Technol. **59**(1), 344–355 (2010)
6. Hull, B., Bychkovsky, V., Zhang, Y., Chen, K., Goraczko, M., Miu, A., Shih, E., Balakrishnan, H., Madden, S.: CarTel: a distributed mobile sensor computing system. In: Proceedings of ACM SenSys, pp. 125–138 (2006)
7. Mohan, P., Padmanabhan, V.N., Ramjee, R.: Nericell: rich monitoring of road and traffic conditions using mobile smartphones. In: International Conference on Embedded Networked Sensor Systems, pp. 323–336 (2008)
8. Liu, J., Wan, J., Wang, Q., Deng, P., Zhou, K., Qiao, Y.: A survey on position-based routing for vehicular ad hoc networks. Telecommun. Syst. **62**(1), 15–30 (2016)
9. Lee, U., Zhou, B., Gerla, M., Magistretti, E., Bellavista, P., Corradi, A.: Mobeyes: smart mobs for urban monitoring with a vehicular sensor network. IEEE Wirel. Commun. **13**(5), 52–57 (2006)
10. Lee, U., Magistretti, E., Gerla, M., Bellavista, P., Corradi, A.: Dissemination and harvesting of urban data using vehicular sensing platforms. IEEE Trans. Veh. Technol. **58**(2), 882–901 (2009)
11. Jiang, D., Taliwal, V., Meier, A., Holfelder, W., Herrtwich, R.: Design of 5.9 GHz DSRC-based vehicular safety communication. IEEE Wirel. Commun. **13**(5), 36–43 (2006)
12. Frenkiel, R.H., Badrinath, B.R., Borres, J., Yates, R.D.: The infostations challenge: balancing cost and ubiquity in delivering wireless data. IEEE Pers. Commun. **7**(2), 66–71 (2000)
13. Sharef, B.T., Alsaqour, R.A., Ismail, M.: Vehicular communication ad hoc routing protocols: a survey. J. Netw. Comput. Appl. **40**(1), 363–396 (2014)
14. Rudack, M., Meincke, M., Lott, M.: On the dynamics of ad hoc networks for inter vehicle communications (IVC). In: International Conference on Wireless Networks (2002)
15. Song, C., Wu, J., Liu, M.: On characterization of the traffic hole problem in vehicular ad-hoc networks. In: Proceedings of IEEE GlobeCom (2013)

16. Perkins, C.E., Royer, E.M.: Ad-hoc on-demand distance vector routing. In: Proceedings of Workshop on Mobile Computing Systems and Applications, pp. 90–100 (1999)
17. Karp, B., Kung, H.T.: GPSR: greedy perimeter stateless routing for wireless networks. In: International Conference on Mobile Computing and Networking, pp. 243–254 (2000)
18. Zhao, J., Cao, G.: VADD: vehicle-assisted data delivery in vehicular ad hoc networks. In: Proceedings of IEEE INFOCOM, pp. 1–12 (2008)
19. Song, T., Xia, W., Song, T., Shen, L.: A cluster-based directional routing protocol in VANET. In: IEEE International Conference on Communication Technology, pp. 1172–1175 (2010)
20. Zeadally, S., Hunt, R., Chen, Y.S., Irwin, A., Hassan, A.: Vehicular ad hoc networks (VANETs): status, results, and challenges. Telecommun. Syst. **50**(4), 217–241 (2012)
21. Korkmaz, G., Ekici, E., Ozguner, F., Ozguner, U.: Urban multi-hop broadcast protocol for inter-vehicle communication systems. In: ACM International Workshop on Vehicular Ad Hoc Networks, pp. 2062–2063 (2004)
22. Song, C., Liu, M., Wen, Y., Cao, J., Chen, G.: Buffer and switch: an efficient road-to-road routing scheme for VANETs. In: International Conference on Mobile Ad-Hoc and Sensor Networks, pp. 310–317 (2011)
23. Ding, Y., Wang, C., Xiao, L.: A static-node assisted adaptive routing protocol in vehicular networks. In: International Workshop on Vehicular Ad Hoc Networks, pp. 59–68 (2007)
24. Song, C., Wu, J., Yang, W.-S., Liu, M., Jawhar, I., Mohamed, N.: Exploiting opportunities in V2V transmissions with RSU-assisted backward delivery. In: INFOCOM 2017 6th IEEE International Workshop on Mission-Oriented Wireless Sensor and Cyber-Physical System Networking (MiSeNet'17) (2017)
25. Johnson, D.B., Maltz, D.A.: Dynamic source routing in ad hoc wireless networks. In: Mobile Computing, pp. 153–181 (1996)
26. Naumov, V., Gross, T.R.: Connectivity-aware routing (CAR) in vehicular ad-hoc networks. In: Proceeding of IEEE International Conference on Computer Communications, pp. 1919–1927 (2007)
27. Lee, K.C., Le, M., Harri, J., Gerla, M.: LOUVRE: landmark overlays for urban vehicular routing environments. In: Proceeding of IEEE Vehicular Technology Conference, pp. 1–5 (2008)
28. Pelusi, L., Passarella, A., Conti, M.: Opportunistic networking: data forwarding in disconnected mobile ad hoc networks. IEEE Commun. Mag. **44**(11), 134–141 (2006)
29. Jain, S., Fall, K., Patra, R.: Routing in a delay tolerant network. ACM SIGCOMM Comput. Commun. Rev. **34**(4), 145–158 (2004)
30. Vahdat, A., Becker, D.: Epidemic routing for partially-connected ad hoc networks, Master Thesis
31. Rashid, B., Rehmani, M.H.: Applications of wireless sensor networks for urban areas. J. Netw. Comput. Appl. **60**, 192–219 (2016)
32. Kafi, M.A., Challal, Y., Djenouri, D., Doudou, M., Bouabdallah, A., Badache, N.: A study of wireless sensor networks for urban traffic monitoring: applications and architectures. Procedia Comput. Sci. **19**(Complete), 617–626 (2013)
33. Piran, M.J., Murthy, G.R., Babu, G.P.: Vehicular ad hoc and sensor networks; principles and challenges. Int. J. Ad Hoc Sens. Ubiquitous Comput. **2**(2), 38–49 (2011)
34. Mednis, A., Elsts, A., Selavo, L.: Embedded solution for road condition monitoring using vehicular sensor networks. In: International Conference on Application of Information and Communication Technologies, pp. 1–5 (2012)
35. Jakubiak, J., Koucheryavy, Y.: State of the art and research challenges for VANETs. In: Consumer Communications and Networking Conference, pp. 912–916 (2008)
36. Wischhof, L., Ebner, A., Rohling, H.: Information dissemination in self-organizing intervehicle networks. IEEE Trans. Intell. Transp. Syst. **6**(1), 90–101 (2005)
37. Al-Sultan, S., Al-Doori, M.M., Al-Bayatti, A.H., Zedan, H.: A comprehensive survey on vehicular ad hoc network. J. Netw. Comput. Appl. **37**(1), 380–392 (2014)

38. Rawashdeh, Z.Y., Mahmud, S.M.: Intersection collision avoidance system architecture. In: Conference Record—IAS Annual Meeting. IEEE Industry Applications Society, pp. 493–494 (2008)
39. Song, C., Wu, J., Liu, M., Gong, H., Gou, B.: RESen: sensing and evaluating the riding experience based on crowdsourcing by smart phones. In: International Conference on Mobile Ad-Hoc and Sensor Networks, pp. 147–152 (2012)

Part IV
Wearable Computing

An Overview of Wearable Computing

Gary M. Weiss and Md. Zakirul Alam Bhuiyan

Abstract This chapter provides a high-level user-oriented overview of wearable computing. It begins by defining wearable computing and describing its key characteristics. It then provides an historical view of wearable computing devices, beginning with an abacus ring from the 17th century and progressing to modern wearable computing devices. Lessons learned from this history and their implications for future wearable devices are discussed. Key application areas for wearable computing are then introduced, and sample applications from each area are described. Two case studies are provided to demonstrate the role of data analytic methods, and how they can yield more powerful wearable applications. The chapter concludes with a summary of the current state of wearable technology.

1 Introduction

Wearable computing has been around in a limited form for several centuries, but has entered the mainstream only recently, due to the increasing availability of very small and low-powered computational devices and sensors. Some of these devices, such as the smartphone, are not perfect examples of wearable computing due to the significant amount of time and attention it takes to access them, but are incredibly important due to their enormous impact on our economy and society. Better examples of wearable computing devices, like smartwatches, have thus far enjoyed only moderate commercial success, while much more ambitious and potentially revolutionary wearable computing devices, like Google Glass, have encountered difficulty with widespread adoption. Although resistance to these new technologies makes the market for such devices uncertain, the future for wearable computing is still quite promising, due to the increasing availability of inexpensive sensors, low-cost processors, and inex-

G. M. Weiss (✉) · Md. Z. A. Bhuiyan
Department of Computer & Information Science, Fordham University, Bronx, NY, USA
e-mail: gaweiss@fordham.edu

Md. Z. A. Bhuiyan
e-mail: mbhuiyan3@fordham.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_10

pensive memory. This chapter will provide the reader with an overview of wearable computing: what it is, how it has developed, what has been created, what applications it can support, and what the future may hold. This overview is not intended to provide a detailed engineering analysis of wearable computing devices and how they function.

We start by providing an understanding of wearable computing. The concept is relatively straightforward since it generally concerns computing devices that can be worn. However, the existing formal definitions and descriptions of wearable computing tend to refine the concept by emphasizing or deemphasizing certain key characteristics of the devices. An early definition of wearable computing was provided in a July 1996 U.S. Defense Advanced Research Projects Agency workshop on "Wearables in 2005," which aimed to predict the future of wearable ten years into the future. This DARPA workshop defined wearable computing as "data gathering and disseminating devices which enable the user to operate more efficiently. These devices are carried or worn by the user during normal execution of his/her tasks" [1]. The key element in this definition is that wearable computing should be used in a natural, unobtrusive, manner. Steve Mann, an early pioneer in the field who created a personal imaging device with a camera and display built within an ordinary pair of sunglasses, described three key characteristics of what he referred to as "WearComp" [2]. According to Mann, a wearable computer is worn, not carried, in such a way as it can be regarded as being part of the user; it is user controllable, not necessarily involving conscious thought or effort; and it operates in real time and is always on.

Common modern examples of commercial wearable computing devices include: wearable fitness trackers like FitBit, smartwatches such as the Apple Watch and Motorola Moto 360, and wearable cameras like the ones popularized by GoPro. Google glass, which is currently available only for development purposes, is the most ambitious commercial wearable computing product to date. It provides general computer functionality to the user via a tiny projection in front of the user's eye, utilizes voice recognition technology, and includes many of the functions of the very early head-mounted displays. The single most popular wearable computing device is the smartphone, although this is not an ideal example of wearable computing given the definitions just provided—the smartphone often needs to be held in one's hand rather than worn, which means that its use is often obtrusive. However, because it is often "worn" in one's pocket, and in some cases can be controlled by a smartwatch, it is a valid wearable computing device and can sometimes even be used unobtrusively.

2 The History of Wearable Computing

This section provides an overview of the history of wearable computing. This will provide insight into the development path of wearable computing, highlight challenges and issues that can occur, and will allow us to identify several patterns—and lessons—that can help us predict the future path of wearable computing.

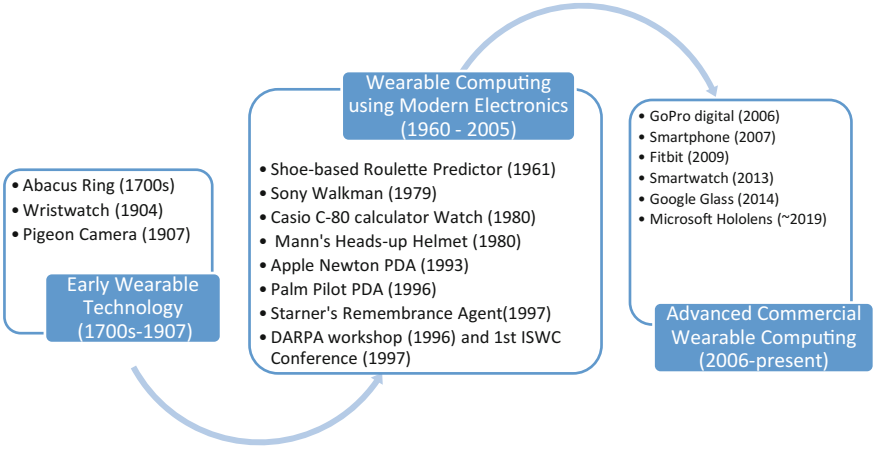


Fig. 1 Timeline of wearable computing products

Interestingly, the history of wearable computing is quite extensive, and even predates the invention of the modern electronic computer. Given the number of wearable devices that have been developed it cannot be comprehensive, so this section instead focuses on breadth—covering wearables from different time periods—and also on the most important and well-known examples of wearable computing. We divide the history of wearable computing devices into three main time periods: Early Wearable Technology (1700–1907s), Wearable Computing using Modern Electronics (1960–2005), and Advanced Commercial Wearable Computing (2006—present). The main distinguishing characteristics between the first and second time periods is the type of technology used, while the main differences between the second and third periods is the level of refinement and commercial success. An overview of all of the products described in this section is provided in Fig. 1.

2.1 Early Wearable Technology (1700–1907s)

The earliest wearable technology had limited computing capabilities and may not have had any electronic components, but nonetheless shared many characteristics with modern wearables. Part of the significance of these early devices is that they were the precursors to more advanced computational devices that were introduced many years later.

2.1.1 Abacus Ring (1700s)

A very early example of wearable computing is the abacus ring from China's Qing Dynasty, which was used by traders in the seventeenth century. The tiny abacus ring was 1.2 cm-long and 0.7 cm-wide, and contained nine wires which each had seven beads. The beads were so small that they could not easily be manipulated by human fingers and instead were manipulated by small pins (it is believed that the rings were often used by women who could use their hairpins for this task). The ring allowed large sums to be quickly tabulated, and hence is an excellent example of non-electronic wearable computing. The abacus ring may be functionally considered a forerunner of or the calculator watches of the 1980s.

2.1.2 Wristwatch (1904)

Another key development in wearable technology had to do with tracking time. As the technology used for making timepieces miniaturized, pocket watches were invented so one could always have a mobile (i.e., portable) timepiece. But removing one's pocket watch to check the time could interfere with important activities, especially in the military, where the synchronization of military maneuvers or actions was often critical. Toward the end of the nineteenth century watches had become sufficiently miniaturized that a German artillery officer was able to strap a pocket watch to his wrist. The first "true" wristwatch was created in 1904 when Alberto Santos-Dumont, an experimenter in heavier-than-air flying machines, commissioned the famous jeweler Louis Cartier to manufacture a small timepiece with a wristband, so that he could check the time on his "wristwatch" while keeping his hands available for flying [3]. While this device did not have any true computing power, it serves as the precursor of the modern smartwatch.

2.1.3 Pigeon Camera (1907)

Around the same time as the invention of the wristwatch, Dr. Julius Neubronner invented a tiny camera, with a timing mechanism, capable of taking a single image (see Fig. 2). The camera was strapped to a pigeon and the resulting aerial images garnered Dr. Neubronner quite a bit of fame. Many of the aerial photographs were turned into postcards. This "pigeon camera" is superficially similar to the popular GoPro camera, because both are "worn." However, there are significant differences, in that the main goal of the GoPro is to take first-person action shots and videos, whereas the pigeon camera was largely designed in order to make it possible to take inexpensive aerial photographs.

Fig. 2 Dr. Neubronner with his pigeon camera

2.2 Wearable Computing Using Modern Electronics (1960–2005)

Electronic programmable computers were introduced in the 1940s, and just a few decades later began to be incorporated into wearable devices. These computers were relatively primitive toward the beginning of this period, but were quite powerful by the end of the period.

2.2.1 Shoe-Based Roulette Predictor (1961)

In 1955 Edward Thorpe conceived of what he considered to be the first wearable computer—a shoe-based device for predicting the outcome of roulette spin. Eventually, in 1960 and 1961, Thorpe built the device with assistance from Claude Shannon, and it was field tested in Las Vegas in 1961 [4]. The device was quite successful and it is expected gain of +44%, determined under laboratory conditions, was validated in Las Vegas and yielded a $10,000 profit. A shoe-based device was used for inputting timing information related to the ball position via toe switches. Based on the physical models that were programmed into the computer, the device would predict which of the eight octants of the roulette wheel that the ball would land in. Two people and devices were involved: one to input the data into one device, while the other would wirelessly receive the prediction via another shoe-based device and place the bet. Thorpe and Shannon kept their invention secret until 1966. The invention was later commercialized by Eudaemonic Enterprises but never generated a huge win [5].

2.2.2 Sony Walkman (1979)

One of the trends that began in this period was the miniaturization of electronic devices. One of the most striking success stories was the Sony Walkman, a 14 oz portable cassette player that was introduced on July 1, 1979. While this was not a computing device—or initially even a digital device—it is very important in the history of wearable computing for two reasons. The first reason is that it was the first electronic device that demonstrated the immense market potential for electronic wearables. More than 400 million Sony portable music players have been sold, and of these 200 million were of the original cassette variety. Thus it anticipated and perhaps suggested the future success of more advanced wearables like the iPhone. The second reason is that the initial Sony Walkman is the predecessor of the portable digital music player, which is a legitimate wearable computing device. While digital music players such as the Apple iPod provided only limited computing capabilities, they ultimately hastened the advent of the smartphone, which included many advanced computing capabilities.

2.2.3 Calculator and Databank Watches (1980)

Simple commercial wearable computing devices began to enter the marketplace in the late 1970s in the form of calculator watches manufactured by Pulsar and HP. However they did not become very popular until the 1980s, when other manufacturers, such as Casio, released their own products. Casio released the C-80 watch with a calculator function in 1980, the T-1500 with a dictionary function in 1982, and the CD-40 with a "databank" function in 1984. The databank could save and recall 10 groups of 16 letters or numerals, eliminating the need to carry a personal phone-number organizer. This watch highlighted the concept of an information device on the wrist. The CD-40 became a major hit, selling six million units in the 5 years after its release [6]. These devices are the predecessors of the modern smartwatch (Fig. 3).

2.2.4 Personal Digital Assistant (1993)

In the early 1990s a new type of commercial portable digital device came out—the personal digital assistant (PDA). These devices included a phone and address book, a calendar, and typically had the ability to take notes using a stylus. These devices also often had handwriting recognition capabilities. The combination of the stylus and handwriting recognition led to a change in user interface and promoted the use of handwriting over keyboarding. Notable PDAs included the Apple Newton, which was introduced in 1993, and Palm Computing's Pilot, which was introduced in 1996. Similar to modern smartphones, these devices were not perfect examples of wearable computing since they normally had to be held in one's hand, but the services that they provided are services that one would want a wearable to provide. Ultimately the functions provided by the PDA were incorporated into the smartphone and hands-

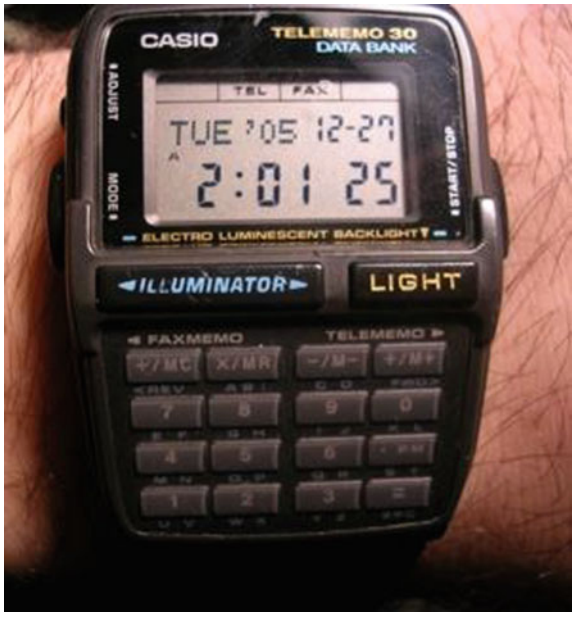


Fig. 3 A casio data bank watch

free wearables like smartwatches that could utilize voice recognition. The use of the stylus—perhaps combined with handwriting recognition—has been almost entirely eliminated from most modern wearable computing devices.

2.2.5 Head-Mounted Displays and Starner's Remembrance Agent (1968–1997)

Much of the most interesting and ambitious work on wearable computing in the late 1960s and the 1970s involved head-mounted displays. The goal for these head-mounted devices was to provide the benefits of traditional computing (e.g., email), but unobtrusively and in a mobile setting, and to also provide new capabilities that are now commonly associated with augmented reality. Ivan Sutherland was involved in very early work in this area, and his head-mounted displays employed partially reflective mirrors to let the wearer see a virtual world superimposed on reality. However, due to the technological limitations of the time, his head-mounted displays had to be tethered in order to obtain the necessary power and computing resources [7]. In 1980, Steve Mann, while still in high school, developed a tetherless helmet with a built-in CRT screen, which allowed one to monitor the screen while walking around [8]. But this system suffered due to the hassle of wearing cumbersome head gear, low-resolution video, eye fatigue, and the requirement for dim lighting conditions. However, over a 16 year span, advances in miniaturization allowed Mann to address many of these issues, and led to the first eyeglass-mounted versions in the late 1980s

and, by 1996, an eyeglass version that was virtually indistinguishable from a normal pair of glasses [8]. Computer control was accomplished via a handheld device, which limited the user's ability to interact normally with his environment. Mann envisioned applications such as capturing visual images to form a pictorial diary of one's day, and computationally augmenting the projected images so that the device operates as a personal virtual assistant for the visually challenged.

Research at MIT Media Lab, led by Thad Starner in the 1990s, focused on using head-mounted display technology to function as a personal assistant via augmented reality [9]. The wearable computer would monitor what the user was doing and then would provide relevant information. This information agent was known as the Remembrance Agent [10]. The head-mounted system used finger tracking to replace a computer mouse, an important advance in the area of wearable computing. The system experimented with a variety of important concepts that are still relevant today: providing textual descriptions for physical objects ("physically-based hypertext"), 3D graphics overlaid on physical objects (e.g., for repair instructions), and even using face recognition to identify nearby people. The authors envisioned a prediction component that would ultimately anticipate the user's needs and act accordingly. This system, and systems like it, highlighted the potential value of augmented reality. Google Glass is the descendent of these various systems.

2.2.6 Formation of Dedicated Research Communities (1996)

By the mid-1990s, there was sufficient interest in wearable computing that the main participants could organize and form their own research community. This led to the Defense Advanced Research Projects Agency 1996 workshop on "Wearables in 2005," which attempted to predict the future of wearable computing [1]. The DARPA workshop was attended by representatives from academia, industry, and the military. It was followed a year later by the First International Symposium on Wearable Computers [11], a conference that is still active today. As mentioned earlier in this chapter, papers presented at these meetings helped to define wearable computing and their defining characteristics. Since this 1996 meeting, there has been a research community continuously dedicated to wearable computing, as well as a series of conferences and workshops focused on the topic.

2.3 Advanced Wearable Computing (2006–Present)

The period that we refer to as advanced wearable computing is largely characterized by highly refined products, designed and engineered for the mass market, which typically achieve enormous commercial success. Some of the products, however, are not particularly ambitious in that they are only intended to perform one or a few tasks very well. Only toward the end of this period do we see some products that rival the

ambitions of the heads-up displays from the 1980s—and these products have thus far largely failed to catch on (e.g., Google Glass), or are not quite ready for market (Microsoft Hololens).

2.3.1 The GoPro Digital Camera (2006) and Other Lifelogging Wearables

One of the early successes of this period was the GoPro camera. While the first version was introduced in 2004, the first digital version was not introduced until 2006 (that version included the ability to take short videos). As the technology advanced, more storage was available and longer videos could be taken. The distinctive aspect of the GoPro devices was that they were designed to be worn in order to take "first-person" pictures and videos while engaged in active sports, such as skiing or surfing. While these devices performed only limited computing functions, their imaging capabilities were similar to those provided for by the early head-mounted wearables—both were capable of capturing first-person video to help record your life.

More recently, a number of very small wearables have entered the market to help continuously record your life. The Perfect Memory Smart Pro Camera from General Streaming Systems, LLC, is a very small wearable that can be clipped on to your clothes that provides continuous recording [12]. Due to storage limitations not everything can be kept and therefore the product's main goal is to allow you to save interesting events that happen to occur. If something noteworthy occurs, you can manually scroll back and save the footage, or tap the device and it will automatically save the last 5 min of footage. Other lifelogging cameras include the Narrative Clip 2, YoCam, Sony Experia Eye, and ION SnapCam [13]. Most of these devices can use BlueTooth to connect to a smartphone to save the images. The notion of lifelogging has not yet caught on, but could become more attractive and desirable with continued improvements in computer technology.

2.3.2 Smartphones (2007)

The most notable commercial success of this time period is the smartphone. The Apple iPhone and the first Android phone were released in the United States in 2007 and 2008, respectively. These multi-function devices incorporated a large number of services and capabilities, including: phone, camera/video recording, Internet connected web browsing, email, phone/address book, digital music player, GPS-based directions, and video game player. Thus they incorporated all of the functions of cell phones, personal digital assistants, and digital music players. They also were able to serve as cameras, portable gaming systems, and GPS-enable mapping applications. Aside from all of these capabilities, smartphones are also miniature computers with significant processing power and reasonable storage capacity.

The smartphone is not an ideal wearable computing device since many of its functions require that it be retrieved from one's pocket or purse and held in one's

hand. The handheld manual operation of the smartphone can interfere with normal daily activities. Thad Starner, who has been wearing heads-up displays for more than two decades, says that it takes about 20s to retrieve a smartphone, and that this delay between intention and action is significant and will reduce smartphone usage [9]. However, the smartphone increasingly serves important roles that do not require it to be held in one's hand. This is because it can connect to other devices wirelessly via Bluetooth, so that its functions can be accessed even when "worn" in one's pocket. Furthermore, it can also serve as a communication hub, providing Internet access to other wearable computing devices. In this sense, the smartphone is currently the most important and ubiquitous wearable computing device available, and at the current time is likely to serve as the central component—for both computing resources and internet access—for other wearable computing devices.

One of the key characteristics of smartphones is that they contain many sensors. Since users often carry their phones on their bodies, this provides the potential for continuous sensing. As smartphones became more technologically advanced, more and more sensors were added. Virtually all smartphones now contain an accelerometer, gyroscope, location sensor (e.g., GPS), light sensor, and magnetometer; some smartphone models also include a barometer, heart rate sensor, and even a dedicated pedometer sensor. The accelerometer is central to many health and fitness applications and allows the smartphone to act as a fitness tracker and step counter.

2.3.3 Fitbit (2009)

The first Fitbit activity tracker was released in 2009, and it accelerated the use of wearables in the health and fitness market. The most basic function of a Fitbit is its pedometer function, and its ability to calculate the distance walked and estimate total calories burned. This basic function is incorporated into the many dozens of Fitbit models. Over time, new models were introduced, which featured sleeker designs and additional functions. The majority of Fitbit products is worn on the wrist or are clipped to one's clothing. The more recent models can connect to your smartphone or computer to upload and analyze data, and there are a variety of social networking features to help motivate the user to become more active. More advanced models can be used to track the duration and quality of your sleep, and you can also use the Fitbit app to log your meals and track your weight. While a smartphone stored in one's pocket can provide many of the same functions via the phone's accelerometer [14], the ease of use and low cost of the activity trackers have proven quite attractive to consumers. Fitbit has been quite successful and experienced rapid growth between 2010 and 2015, when its revenue increased from $5 Million to $1.8 Billion. However, it experienced significant problems in 2016 due to manufacturing problems and unexciting product upgrades, and now is facing competition for Apple and others as consumers gain interest in more sophisticated wearables, like smartwatches, which can provide the same capabilities, as well as additional capabilities.

2.3.4 Smartwatches (2013)

The next major commercial advancement in wearable computing was the introduction of the modern smartwatch. Many earlier digital watches could claim be to "smart" in one sense or the other, but the modern smartwatch did not truly arrive until the introduction of the Pebble in July 2013. The Pebble was funded by an enormously successful Kickstarter campaign, which raised $10.3 Million. By 2014 Pebble sold its one-millionth watch, but it shut down its operations by 2016 due to the flood of more technologically advanced watches entering the market. Some of the notable smartwatches that superseded the Pebble include the Samsung Galaxy Gear (2013), the Motorola Moto 360 (2014), the LG G-Watch (2014), and the Apple Watch (2015). These watches generally could only provide full functionality when paired with a smartwatch, which provided long-range data communication access, including access to the Internet.

Smartwatches have many useful features and can support a variety of applications. One of the key functions of a smartwatch is that it provides improved accessibility to one's smartphone, by eliminating the delay associated with removing a smartphone from one's pocket or purse. By using the smartwatches wireless connection to one's smartphone, one can access common smartphone applications within a few seconds, via simple graphical menus provided by the smartwatch screen, or via voice control. With this capability one can control music play, get travel directions, or send a text message. By providing this improved interface a smartwatch removes many of the barriers that prevented the smartphone from being a true wearable computing device.

Smartwatches also can provide some functions beyond what is provided by smartphones, due to their placement on the body. While smartwatches do not yet contain all of the sensors present on a smartphone, they typically provide an accelerometer and gyroscope, and often include a heart rate monitor. Thus smartwatches can provide pedometer functionality independent of a smartphone and can also recognize more sophisticated activities, including hand-based activities [15]. Smartwatches are especially recognized for their health and fitness applications.

2.3.5 Google Glass (2014)

Google Glass is an optical head-mounted display designed to look like a pair of eyeglasses, but with both of the lenses removed. The user can view the projected image by looking up, as the viewable projection is not directly in front of the eye, so it is less intrusive and does not impair human-to-human interaction. This is perhaps the most ambitious commercial wearable computing product that has ever been released. A prototype of Google Glass started selling in the United States on April 15, 2013 for $1,500, and became available to the general public on May 15, 2014. The product was pulled from the market on January 15, 2015, due to many criticisms about the design. It will not be rereleased until it has been significantly improved, but right now it appears that any new release will first be aimed at industrial users. A number of industrial applications are currently under development.

Google Glass has a number of important features. First, like previous head-mounted displays, it has a camera capable of taking first-person pictures and video, although the limited battery life precludes continuous video recording. It also contains a touchpad, which is located on one outside edge of the frame. The device can also be controlled via voice commands. Responses from the product can be relayed to the user visually or can be relayed via audio using bone conduction through a transducer that sits beside the ear; this setup means that others who happen to be nearby will not be able to hear the audio responses.

Google Glass has the potential to support an enormous number of applications. Navigation is one natural application, and an early Google commercial shows how such an application can help one navigate a city, and how relevant information (e.g., subway information) can be automatically displayed in a context-sensitive way. Google Glass can also be used to provide instructions while assembling a new piece of furniture or following a new recipe, or set a reminder or calendar entry with just a voice command. It can also be used to send texts or email with voice commands, or even make a video call. A smartphone can do many of these functions but, as mentioned earlier, the time it takes to physically access the smartphone is a stumbling block for many tasks—and holding the phone interferes with performing other tasks. There is also an expectation that industry-specific applications will be developed and that Google Glass will be used extensively in industrial settings. For example, an employee in a warehouse could receive a notification of which product to retrieve, and then Google Glass could navigate the employee to the proper location.

Google Glass raised many privacy concerns, which were well publicized and disseminated by the mass media. These concerns are important since they are a barrier to adoption for both Google Glass and potentially other similar future wearables. One concern involved the ability to take pictures of others without their permission and knowledge. This led to some bars and other establishments banning Google Glass [16], and to suggestions that the product not be worn inside of public bathrooms. There was also great concern that facial recognition applications could automatically identify strangers and then display information about them from the Internet. Some experts feel that ultimately people will come to accept the technology and ignore the privacy concerns, just as has happened with other new technologies—but others are not so sure. Because Google Glass sales were quite limited prior to the suspension of sales, the ultimate impact of such concerns is still unknown.

2.3.6 Microsoft Hololens (Estimated 2019)

Hololens is a pair of mixed-reality glasses, developed by Microsoft, which projects 3D objects ("holograms") into the user's environment. The current product is not ready for general usage and is intended for developers; there is no product release date but various estimates place a commercial release for 2019. The Hololens is quite different from Google Glass and serves a very different purpose—although there is some overlap. First, the Hololens is much larger than Google Glass and, although tetherless, is primarily intended to be used in a fixed environment. That is because

while Google Glass is a perfect example of augmented reality, where the emphasis is on overlaying information on top of the real world, the Hololens is for *mixed* reality, where the emphasis is on the computer generated 3D object(s). The Hololens is not primarily considered a virtual reality system because the 3D image is only viewable in a relatively small section in the center of one's vision—it is designed to allow the user to work in the real world. A sample Hololens application would allow a student to interact with a 3D image of a human heart, show how to trouble-shoot a printer jam, or allow a user to take a tour of a famous site. The Hololens includes 3D sound speakers and can interact via spoken commands or gestures.

2.4 Lessons Learned from the History of Wearable Computing

The history of wearable computing provides many lessons, which can also tell us something about how wearables will continue to evolve in the future. The key lessons are enumerated below, and then discussed and justified in subsequent subsections. Lesson 5 is the only lesson that cannot be fully justified at the present time.

Lesson 1 Wearable computing devices should fulfill genuine needs.
Lesson 2 Wearable computing devices are most successful when they satisfy multiple needs.
Lesson 3 Wearable computing devices should be very quick to access and use—and this should be supported by the device's user interface and placement on the body.
Lesson 4 Wearable computing devices should be "always on" and available.
Lesson 5 Wearable computing devices should preserve the privacy of the user and bystanders.

2.4.1 Wearable Computing Devices Should Address Genuine Needs

This is perhaps the most basic lesson and does not require much analysis: wearable computing devices should satisfy real needs of the user. Although this lesson seems trivial, it is not given current consumer perception of wearables: many consumers find wearables to be useless—as well as unattractive and expensive. Most successful wearable devices have a corresponding non-wearable analog, because the need existed prior to the technological advances that enabled a wearable version of the device. As was discussed earlier in this section, the need to track time led to large timepieces (clocks), which were subsequently replaced by their non-digital miniaturized counterparts (wristwatches), which were subsequently replaced by digital watches and then smartwatches. The need to perform calculations led to the abacus, which led to the smaller abacus ring, which then led to the calculator watch—and subsequently to the smartwatch. The need to play recorded music led to the record

player, which led to the smaller Sony Walkman (an analog version followed shortly be a digital version), which led to the iPod, and then the iPhone. One way to predict future wearables is then to identify devices that satisfy a need, but cannot be turned into wearables due to technological constraints. Video-based lifelogging is one example of an application that may not yet be feasible, but which may become possible in a few years (i.e., a future version of Google Glass or a competing product may be able to take continuous videos). Similarly, wearable technology that provides high quality monitoring of multiple medical conditions, using sensors deployed over the body (perhaps embedded in clothing), may also become possible over the next decade.

2.4.2 Wearable Computing Devices Are Most Successful When They Satisfy Multiple Needs

Wearable computing devices are often initially developed to satisfy one need—or a narrow range of needs—but history shows us that over time they are often merged into a single device. This occurs even if the merged device does not perform quite as well at satisfying each individual need. Apple Inc. provides perhaps the best example of how a successful wearable device is supplanted by a more powerful, and general, device. For many years Apple produced an incredibly successful series of digital music players, which included the iPod classic, iPod Mini, iPod Nano, and iPod Touch. For a period of several years these music players generated between $5B and $10B in revenue for Apple, with peak worldwide sales of 54 million devices in 2008 and 2009. However, these sales were eventually cannibalized by the introduction of the iPhone and Android smartphones. The smartphones virtually eliminated the market for standalone digital music players. Smartphones also largely replaced other single-purpose wearable computing devices, such as personal digital assistants (PDAs) and handheld GPS trackers; it also seriously impacted the market for portable game players and digital cameras (consumers still purchase digital cameras, but mainly high quality models with powerful optical zooms). Smartwatches, which support multiple applications, have begun to impact the sale of wearable fitness trackers like Fitbit. Based on past history, the future of fitness trackers is not very bright if smartwatches are able to effectively satisfy several user needs—and thus become ubiquitous. But at this moment in time, smartwatches have not yet achieved this status.

Consumers clearly want the convenience of wearable devices that can handle multiple tasks. The merging of these devices yields many benefits. Cost savings is one benefit. The reduction in the number of devices that need frequent charging is another benefit. Each device also places some burden on the user to carry it around, and thus the merging of devices can reduce this burden. The need to merge these devices may be reduced in the future if they can connect wirelessly to share common services (many tasks require the ability to send and receive data). Thus, in the future we may see a reversal of this trend if power requirements drop sufficiently, so that some wearables can be very small and yet communicate with more powerful wearables

(e.g., smartphone) via Bluetooth or a body network. But devices will only separate if there is some concrete benefit (e.g., it achieves a more convenient body location).

2.4.3 Wearable Computing Devices Should Be Very Quick to Access and Use—And This Should Be Supported by the Device's User Interface and Placement on the Body

Wearable computing devices should be unobtrusive. That requires a user interface that permits for quick input, while minimizing any loss of focus by the user, and convenient location on the body. The progression seen in many wearables demonstrates this principle. Counting and calculation devices progressed from the abacus ring, which was hard to use since it required a pin to move small beads, to a digital calculator watch with small buttons, to a smartwatch that employs a graphical interface and can also operate via voice recognition. The personal digital assistant also provides a lesson. The PDA relied on a stylus, and partially compensated for this by allowing users to write in cursive, which was optionally converted into printed characters via handwriting recognition. While this interface appeared to be effective and quite advanced at the time, it was cumbersome in that it required the user to first access the stylus—which was both time-consuming and distracting. This interface was subsequently replaced by small physical keyboards, and then, as the PDA functions were subsumed by smartphones, by virtual keyboards on a touchscreen. The voice recognition capabilities of smartphones also represent an improvement in user interface.

Proper location is also important for wearables. Ideally a wearable should be located so that it is easy to access and use. Location was the motivation for the development of a wristwatch and then a smartwatch, as it was important to be able to tell time without going to one's pocket (for a pocketwatch). The smartphone, which is often located in a pocket or purse, is poorly placed for many of its intended purposes, but this is deemed acceptable only because the poor location is counterbalanced by the benefit of combining many devices into a single device. Perhaps the best reason to own a smartwatch right now is not due to the new functions it can provide (e.g., heart rate monitoring), but for its ability to make many smartphone functions available from the user's wrist; thus many smartwatches can be considered extensions of the smartphone. Even Google Glass can be viewed as addressing the location issue, since much of the information that we want is best conveyed visually, and Google Glass places that visual information right in front of our eye. Google Glass can even provide audio information to the user via bone conduction, which has the added benefit that bystanders will not be able to hear the information. While Google Glass did not turn out to be a successful consumer product, it was certainly not due to the convenient location of the wearable.

Based on historical patterns, we can conclude that the user interfaces of wearable computing devices will continue to improve and body placement may also improve as technology makes this more feasible. Voice recognition may be used in more wearable computing devices, even if that means some of them may need to connect to the

smartphone to provide this capability. Applications will migrate to more natural body locations as more wearable computing devices are able to connect to the smartphone for computational and data communication resources. Things like email, which are now typically displayed on the smartphone, will more often be accessed on the wrist via a smartwatch, and then made directly accessible via new commercial wearable devices like Google Glass. Health and wellness applications, currently one of the most popular applications areas for wearable computing, will improve as wireless sensors move to more informative body locations. This will eventually occur as wearable sensors are routinely embedded in our clothing and accessories (e.g., belt, shoes).

2.4.4 Wearable Computing Devices Should Be "Always on" and Available

Wearable computing devices should be on continuously and should always be available. Most of the popular wearable computing devices, such as smartphones and smartwatches, essentially meet these criteria since they can operate for an entire working day. There was, and still is, some resistance to smartwatches, because people are not used to charging a watch every evening, and found the task burdensome. But people are adapting to this need, and some smartwatch manufacturers have responded by designing smartwatches that can operate several days on a single charge.

Continuous operation is still an issue for some wearable computing applications. Lifelogging, which entails logging everything that goes on around you, requires continuous recording capabilities. Devices like the GoPro are not capable of lifelogging, so simpler devices were developed, but these have not yet proven to be popular, and often have significant limitations (e.g., video is deleted within a few minutes if not saved). Lifelogging is not supported by Google Glass simply because the device will run out of power within a few hours of continuous video recording. Even some fitness applications, when run continuously on a smartphone, may drain the phone battery prior to the end of the day. As technology advances, lifelogging, and other power-hungry applications, should become capable of continuous operation. It will also allow smaller devices (with smaller batteries) to operate continuously, and this should result in wearable computing devices being incorporated into clothing and accessories.

2.4.5 Wearable Computing Devices Should Support Privacy

The final lesson is that privacy is a concern and may impact the adoption of wearable computing devices. Wearables can yield more concerns about privacy than traditional computing equipment because they are always with the user and can track highly personal information, such as the user's location. Because wearable computing devices move with the user and hence come in proximity to many other people, there is one privacy concern unique to wearables: they threaten the privacy of non-users. There

were initially some privacy issues with cellphones and smartphones for this very reason, due to their camera and video-recording capabilities. These concerns focused on their presence in bathrooms and locker rooms. While these concerns still exist, they did not prevent the adoption of these devices, and the issue is largely addressed by the social convention that these devices not be used in environments where people may be unclothed. There are also privacy issues concerning the amount of information that wearables collect about the users (location data, health data, etc.) and potential misuse of this information, but thus far this concern has had not substantial impact on the adoption of these devices.

The privacy issue came to the forefront with the initial introduction of Google Glass. There were two specific privacy concerns that received a great deal of media attention. The first was the ability to surreptitiously record others. Given the placement of Google Glass, this is much bigger issue than for a smartphone. The second issue is far more interesting, since it has to do with the ability the seamlessly access, merge, and display information. Google Glass is capable of supporting face recognition, so it would be possible for the device to identify people in a crowd, collect publically available information about them from the Internet (including from social media), and then display that information to the user. This can all take place without anyone other than the user knowing that it occurred. Even though the devices were never deployed widely, there was tremendous resistance, with some bar owners saying they would not permit the devices into their establishments. Given that devices similar to Google Glass can address many user needs, these privacy issues will likely rise again in the future. If Google Glass had remained on the market as a consumer product, we would have a much better idea if the privacy issues would have been prohibitive, or if people would eventually become accustomed to the devices and inured to the privacy concerns. Since this question has not been resolved, it is difficult to quantify the importance of privacy and its impact on the adoption of new wearable technology. The best we can say is that wearables should address the privacy issue as completely as possible, especially if it has little impact on the functioning of the device.

3 Applications of Wearable Computing

Wearable computing can support a wide variety of applications, as demonstrated by the historical overview provided earlier in this chapter. Nonetheless, past and current technology has focused on a few key industries, and the taxonomy presented in Fig. 4 highlights these industries. Some industries, such as Medicine and the Military, are key users of wearable computing technology because of the tremendous costs associated with performing at anything but peak efficiency. Education, which includes training, has a great deal to gain as wearable computing can provide a more personalized and immersive experience than traditional methods. Wearable computing applications are just beginning to be used in many businesses, but this market should explode in the coming years as specific applications are developed

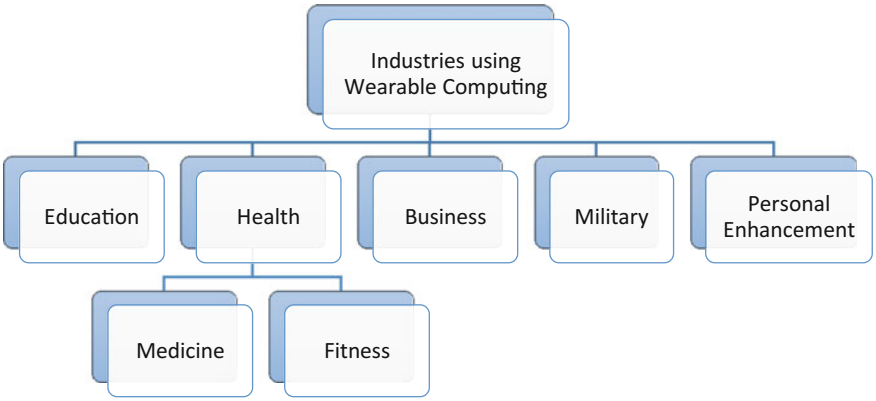


Fig. 4 Major industries employing wearable computing

that enable employees to operate at greater efficiency. The last main entry, "Personal Enhancement," is not what is generally thought of as a major industry, but has been the focus of wearable computing since its inception. This category includes personal assistants and information assistants, and includes any technology that *generally* extends the capabilities of a person that is not tied to a particular industry.

The taxonomy presented in Fig. 4 is not that different than the one provided in a 2014 article on wearable computing applications [17]. That article divided the applications into five categories: health care and medical; fitness and wellness; infotainment; military, and industrial. Most of the applications covered under infotainment are subsumed by the "Education" or "Personal Enhancement" categories.

The industry view is just one way to organize applications of wearable computing. One could also organize them based on the type of capability, or specific technology, that the application provides (the personal enhancement "industry" could fit under this alternative taxonomy scheme). Although the applications described in the remainder of this section are grouped based on industry, it is useful to understand this alternative taxonomy. It is not possible to provide an exhaustive list of the types of capabilities and technologies that wearable computing can provide, but the following list covers the most popular wearable computing capabilities, and also covers all of the applications described in this chapter

- Augmented Reality
- Context Awareness
- Communication and Media
- Sensors and Sensor Mining
- Crowdsourcing
- Social Networking.

Augmented reality provides the user with a view of the real world, but integrates into this view additional information, such as 2D or 3D images, video, text, or audio (this can be contrasted with virtual reality, where the user is placed entirely into a

virtual environment). Many of the early heads-up display units provided some form of augmented reality, but were quite cumbersome to wear and often did not even allow a direct view of the world. More recent systems allow a direct view of the world, are much smaller and lighter, and tend to be worn like eyeglasses—like Google Glass. The goal of augmented reality systems it to assist people as they perform real-world tasks [18], which is an example of intelligence amplification [19]. Augmented reality applications span virtually all industries, and specific examples of these applications are described throughout the rest of this section.

Context-awareness is a general capability that provides an understanding of the context in which the wearable computing device is operating. The context awareness can be utilized by various applications so that they behave more intelligently and are responsive to the environment. Context awareness can be used so a smartphone does not put a call through when the user performing an activity that would preclude a conversation, and can even be used by the computing devices themselves to optimize their resource utilization (i.e., by turning off power-consuming capabilities that the user would not use in the current context). Military applications will generally want to be context aware so that the user can respond appropriately based on what is going on around him.

Many wearables are designed to support communication in all of its many forms, including: phone, email, and text. Communication also includes the capability to access the Internet and to retrieve and play media files, such as music and video. Smartphones and smartwatches are the wearables that currently are most directly tied to communication. A much broader range of wearables utilize communication to share the data that they collect.

Wearable computing devices typically contain many sensors, and those sensors are central to many applications. For example, fitness devices use an accelerometer sensor to count steps, medical wearables use sensors to record patient vital signs, and wearables that support navigation use the GPS sensor to establish location. In many cases the ability to collect sensor data is paired with a data analysis tool, or predictive model, which allows the application to make inferences from the data. Crowdsourcing is a related type of capability, since it normally involves sensor data, but in this case the data is collected from a large pool of subjects. An example application is a navigation system that utilizes crowdsourced traffic data to avoid congestion and minimize total travel time. In such cases the traffic information is determined by analyzing the GPS traces of a large number of automobile drivers. Social networking capabilities also use data, but in this case for social reasons. For example, an application may link you to other people that also jog in your neighborhood. The remainder of this section describes applications of wearable computing to the five industries identified in Fig. 4.

3.1 Education

Education and training is an industry that can benefit from wearable computing and, in particular, from augmented reality. One benefit of augmented reality in an educational setting is that text, graphics, and even video, can be superimposed on the student's or trainee's environment. In one example, a publisher in Tokyo released textbooks that revealed augmented educational content when pages of the book were viewed through a smartphone running an augmented reality app [20]. In another example, a product called AR Circuits (arcircuits.com) allows you to experiment with electrical circuits without purchasing any real hardware. Instead, you print out circuit cards on paper, connect them on a flat surface, and then when you bring them into view of smartphone camera running a special app, the circuits come alive on the smartphone screen as working electrical components. There are currently hundreds of education-related augmented reality apps available that allow students to interactively explore the solar system, the human body, historical sites, and other environments in 3D, with educational information superimposed on the structures. The use of augmented reality in education is expected to grow rapidly in the coming decades.

Medical education, especially anatomy, can especially benefit from augmented reality given the high cost of cadavers. One augmented reality system geared toward undergraduate anatomy education integrates a public CAT scan data set with an actual image of the user, so that that user can effectively "see" inside of his body and navigate through the internal structures of the body with hand gestures [21]. The University of Nebraska is betting on this technology as it is opening a $119 million virtual and augmented reality facility to educate the next generation of healthcare workers.

3.2 Health

The healthcare industry has been one of the early adopters of wearable computing technology. It was adopted in the medical domain due to the need for high performance and the cost of errors, and was adopted in the fitness domain because of the relative ease of developing simple and low-cost fitness tracking applications.

3.2.1 Medicine

Wearable computing provides many benefits in medicine, especially since it permits cost-effective continuous monitoring of vital signs and other data when the patient is not in a hospital or doctor's office. This data can then be used to make health decisions. This area has attracted sufficient attention to have its own term: mobile health, or mHealth [22]. While mHealth is not focused solely on wearables, wear-

ables play a central role in the discipline. Many wearable devices are being developed for mHealth, although as of yet these wearables have not had the mainstream success of the activity trackers used for fitness monitoring. The new devices that are being developed exploit the advances in sensing technology, which permit low-power sensors to monitor the functioning of one's body. These sensors are usually placed on the body, but may be inserted into the body. Several representative medical wearables are described in this section.

The VitalPatch® biosensor, produced by VitalConnect (http://vitalconnect.com), measures heart rate, respiratory rate, skin temperature, and single-lead ECGs, and transmits the data in real time to healthcare providers who can intervene if necessary. It can even detect a fall. The data is analyzed using predictive analytics that can identify problems before they become serious. Data can be streamed and stored in the cloud, and from there be shared with both the patient and doctor. The biosensor-based device has a battery life of four days. The MiniMed® 530G System is a wearable device, comprised of a sensor and insulin pump, that monitors glucose levels and automatically dispenses insulin in way that mirrors that of an actual pancreas [23]. Status information can be relayed to your smartphone via a specialized smartphone app. The Zio cardiac monitor is a patch from iRhythm (iRhythm.com) that can comfortably be worn by a patient for two weeks at a time, and can be used to monitor for cardiac abnormalities. A wearable called Quell (www.quellrelief.com) uses an accelerometer to gauge a user's activity level and adjust its stimulation intensity to alleviate pain. The device uses Bluetooth to connect to a smartphone app, where a user can control the device's features and track therapy and sleep results. Finally, WristOx2 by Nonin Medical (www.nonin.com) is a wristwatch-type device for people with asthma who are at risk for congestive heart failure and chronic obstructive pulmonary disease (COPD). This device monitors a user's heart rate and blood oxygen levels. These devices all are similar in that they rely on accurate biosensors and provide for automatic analysis of the data, although the analysis may occur on the person (i.e., on the device or connected smartphone) or at a remote location that receives the data.

Augmented reality also has applications to medicine, beyond the medical education application mentioned earlier. One study demonstrated that Google Glass has many potential benefits for pediatric surgeons, such as making hands-free photo and video recordings [24]. There is also a medical device that has been in use since 2005 that employs augmented reality to identify a vein by using near-infrared light and then projects a green light onto the skin's surface in order to facilitate intravenous injections [25, 26]. Another medical application comes from a company called Brain Power, which develops Google Glass applications and hardware add-ons, to help improve the life of people with autism. The app will help children with autism to focus on the faces of others by presenting exercises as games and providing points for proper behavior; it will also train its subject to identify emotions based on facial expressions [27]. Another company called VA-ST makes wearable "Smart Specs" for partially sighted or legally blind people, which can assist them in navigating the world [28]. An augmented reality-based surgery system has even been used for advanced laparoscopic liver surgery [29].

3.2.2 Fitness

Wearables have been used for several decades to help determine what activities a person is performing, which can serve as the basis for fitness tracking. The initial motivation for determining a user's activity was to simply understand more about the user and their daily activities, or to allow wearables to be "smart" by acting in a context-sensitive manner. Much of this work was research-based and did not necessarily get incorporated into commercial products. Later work focused mainly on fitness activities, such as walking, and focused on quantifying a user's physical activity. This work led to the many commercial fitness tracking devices. The commercial activity trackers that have been developed tend to focus exclusively on basic fitness activities like walking and jogging, while the research-based systems often also include activities of daily living, such as brushing ones teeth, sitting, reading, typing, etc. A relatively exhaustive list of applications for activity tracking technology is provided by Lockhart et al. [30].

Early research into activity tracking utilized custom sensors that were strapped to various parts of the human body [31]. Over time much of the activity recognition research migrated to commercially available mobile devices, which contain accelerometers. One of the earliest studies to use commercial smartphones for activity recognition showed that a smartphone could identify walking, jogging, stair climbing, sitting, and standing activities [14]. This was about the same time that the original Fitbit activity tracker was released. Smartphones are not ideal for activity recognition because of their inconsistent placement, but smartwatches do not suffer from this problem, and hence they are now viable alternatives to dedicated fitness trackers.

Commercial activity tracking wearables that focus exclusively on fitness tracking and related health applications have been a commercial success, led by sales of Fitbit, which sold over 22 Million fitness trackers in 2016. Fitbit devices act as a pedometer, calculate calories burned, measure progress toward goals, allow users to share their fitness results with others, and may even track sleep. Fitbit now sells watch-based fitness trackers, and virtually all brands of smartwatches now include fitness tracking capabilities, most notably the Apple Watch. The Apple watch will track steps taken and calories burned, but will also tell you if you get up to move regularly, and provide a graph that tells you when throughout the day you were active. The fitness market has so dominated the wearables market that they two terms are sometimes used interchangeably.

3.3 Business and Manufacturing

Business and manufacturing have been using wearable computing for a while, although it has not caught on as quickly as in the health and military communities. Wearable computing can have an especially pronounced benefit in manufacturing and other business activities where the worker needs both hands, but also needs to access complex information. Boeing was one of the first businesses to recognize

this. In the 1990s, Boeing needed hundreds of workers to assemble wiring harnesses for aircraft. This task required both hands, but also reference to voluminous paper instructions. Boeing deployed head-mounted displays, which removed the need for printed instructions, and this led to improvements in worker productivity.

Google Glass, which is currently discontinued as a consumer device, is in active use for manufacturing. AGCO, a company that makes farm equipment, uses Google Glass to assist in the process of assembling tractor engines. A worker can scan the serial number of a part, and relevant manuals, photos, and videos will appear. Voice commands can also be used to bring up more information. Google Glass is a much better mechanism for obtaining information than the tablet computers that were used previously—and often dropped and broken. Google Glass applications are also being used to efficiently guide warehouse workers to the locations of products that need to be retrieved. In one study using a Dutch logistics company, within its first week of use Google Glass led to a 15% increase in stock picking speed and a 12% decrease in worker errors [32]. More conventional uses of wearable computing can also assist warehouse and retail productivity: a host of companies sell ring barcode scanners that permit workers to scan items and thus free up the worker's hands.

Wearable computing can also improve worker safety. The Reflex wearable from Kinetic (http://wearkinetic.com) automatically detects high-risk postures and notifies the worker of the unsafe position. Over time the device teaches the workers to have good biomechanics, and significantly reduces the number of unsafe postures, which leads to reductions in worker injury. Meanwhile, Life by Smartcap (http://smartcaptech.com) detects when a subject is fatigued and in danger of falling asleep, via EEG readings that are automatically captured by the device. This information is transmitted in real time to a central monitoring station, which can take action. It is employed by industries such as mining and construction, where a lack of alertness can cause serious injury or death.

3.4 Military

The military is often an early user of advanced technology and this holds true for wearable computing. Wearable computing can be particularly beneficial by allowing soldiers to focus on what is happening on the battlefield. As an example, if a soldier needs to "look down" to access certain information, like a map, this can put him at risk in hostile situations. Applied Research Associates ARC4 augmented reality system addresses this issue by overlaying tactical information onto a soldier's field of vision [33]. The system can also display the location of teammates, information about geographical features and buildings (including their distance), and keep the soldier on a predetermined route by displaying a waypoint at a short distance. The soldier can even tag points of interest to share with their teammates.

Wearables also have an important role in the military for monitoring the health of soldiers. There is currently a great deal of concern for traumatic brain injury (TMI). The Defense Advanced Research Projects Agency (DARPA) has developed

a wearable blast gauge that measures the impact from an explosion on the solider and automatically notifies medics to respond. The data provides useful information about the severity of the blast and can ensure appropriate treatment. Another issue concerns soldier being pushed beyond their limits, as they are subjected to very high or low temperatures in situations where they must exert themselves. Wearable, chest-based sensors can now determine when soldiers are reaching their physical limits, so they can rest or don protective clothing.

One of the most exciting wearable for the military, which is gaining a great deal of attention lately, is the artificial exoskeleton. The exoskeleton system, known as HULC (Human Universal Load Carrier), allows soldiers to move with less effort, so that they can walk and run for long periods of time without getting tired—even while carrying heavy loads. The units use artificial intelligence to ensure that they are properly amplifying the soldier's intended movements. While these devices can be quite bulky, they meet the definition of wearables provided earlier in the chapter, since they allow the user to operate more efficiently and are used in a natural, unobtrusive, manner.

3.5 Personal Enhancement

A vast number of wearable computing devices aim to generally enhance the effectiveness of the user, without targeting a specific industry. Many examples that fall into this category were described in the historical overview provided earlier in this chapter, and hence will not be repeated here. Perhaps the best examples include the head-mounted displays [8] and Starner's Remembrance Agent [9, 10], which can serve as personal assistants. The personal digital assistants that arrived in 1993 were not nearly as ambitious, but nonetheless supported many of the function of a personal assistant. Smartphones and smartwatches also provide many of the capabilities of a personal assistant, but also provide other capabilities that expand the user's capabilities—from the ability to listen to music to the ability to communicate via email or text. Google Glass, which was described in detail earlier, is capable of providing some of the most advanced applications in this area. In particular, the augmented reality capabilities built into Google Glass extend the capabilities of the human user by seamlessly providing context-appropriate information (e.g., by providing subway information when the user looks at a subway entrance).

4 Case Studies: Activity Recognition and Biometrics

This section describes two research studies related to wearable computing: one involving activity recognition [15] and the other involving biometric identification [34]. These studies only require a commercially available smartphone, which is worn in the pants pocket, and a commercially available smartwatch, which is worn on the

dominant wrist. Both the smartphone and smartwatch contain an accelerometer and gyroscope, and the data from both of these sensors is captured while the user performs a variety of activities. This data is then used to build and evaluate a model to identify the physical activity the user is performing (activity recognition), and to build and evaluate a model to identify or authenticate the user's identity (biometrics). The general approach for both case studies is very similar: training data is captured from the devices and then predictive models are generated using common machine learning classification algorithms.

These two case studies demonstrate how future wearable computing technology can progress to better satisfy the needs of users. The case study on activity recognition shows that today's activity tracking applications are quite primitive in what they can track, and that the technology is capable of tracking a much wider set of activities. This example shows the potential for existing wearable computing applications to become smarter through the use of machine learning and data mining methods. Wearable computing devices tend to capture a tremendous amount of data and it does not yet appear that this data is being fully leveraged. The case study on biometrics also shows how data mining and machine learning methods can better exploit data, but it also demonstrates the potential of wearable computing applications to reduce the burden on its users by automating tasks and making them completely unobtrusive. Currently, computer security is accomplished via the use of passwords, which must be manually entered, or via biometric technology such as fingerprint or face recognition. All of these take effort on the part of the user, whereas the proposed application employs the user's motion data to identify them, and hence can be accomplished without any special effort by the user (this assumes the work that is described is extended to include continuous biometrics).

4.1 Data Collection

The activity recognition and biometric models are generated using supervised learning methods and require labeled motion data. The data is also required to evaluate the models. Data was collected from 51 test subjects, each of whom performed 18 routine activities for 3 minutes each, with an Android smartphone in their pocket and an Android-Wear smartwatch on their dominant wrist. A custom-developed Android application sampled the tri-axial accelerometer and gyroscope sensors on the smartphone and smartwatch at 20 Hz. The raw time-series sensor data, for both the accelerometer and gyroscope, was recorded in the following format:

$$< \text{timestamp}, x, y, z >$$

The timestamp is measured in nanoseconds and the x, y, z values correspond to the three spatial axes. The x, y, and z values are measured in m/s^2 for the accelerometer and in rad/s (radians per second) for the gyroscope. The 18 activities included in the study are listed below, organized logically into three categories.

General Activities (non hand-oriented)

- Walking
- Jogging
- Stairs (ascending and descending)
- Sitting
- Standing
- Kicking a Soccer Ball (two people)

General Activities (hand-oriented)

- Dribbling a Basketball
- Catch with a Tennis Ball (two people, underhand)
- Typing
- Writing
- Clapping
- Brushing Teeth
- Folding Clothes

Eating Activities (hand-oriented)

- Eating Pasta
- Eating Soup
- Eating a Sandwich
- Eating Chips
- Drinking from a Cup.

4.2 Data Transformation

Most classification algorithms cannot directly handle time-series data, but rather expect an unordered set of examples. So that these classification algorithms can be used, the time-series data is transformed into examples via a sliding window approach. A 10-s window is moved over the time-series data, without overlap, and the low-level sensor data in each 10-s segment is represented as a single example via the formation of 43 descriptive, high-level features. The features, which are listed below, are used for both the accelerometer and gyroscope sensor data, and are used for both the activity recognition and biometrics tasks. The value in the square brackets indicates the number of features generated. When three features are generated they correspond to the three spatial axes.

- Average [3]: Average sensor value (each axis)
- Standard Deviation [3]: Standard deviation (each axis)
- Average Absolute Difference [3]: Average absolute difference between the 200 values and the mean of these values (each axis)
- Time Between Peaks [3]: Time between peaks in the sinusoidal waves formed by the data as determined by a simple algorithm (each axis)

- Average Resultant Acceleration [1]: For each of the sensor samples in the window, take the square root of the sum of the square of the x, y, z axis values, and then average them.
- Binned Distribution [30]: The range of values is determined (maximum–minimum), 10 equal-sized bins are formed, and the fraction of the 200 values within each bin is recorded (each axis)

After each example is formed, a label is appended that indicates the activity the participant was performing, and a numerical ID is also added that uniquely identifies the subject.

4.3 Activity Recognition Experiments and Results

The activity recognition task is to identify an activity based on 10 s of sensor data. A classification model is built from a subset of the collected data, the training set, and is subsequently evaluated on a separate subset of the collected data, the test set. Two types of models are induced and evaluated: personal models and impersonal models. Personal models are built for each user, using training data *only* from that user. This requires the user to execute a training phase, which can be inconvenient. Impersonal models, also known as universal models, are built using training data from a panel of *other* users, and requires only a single (universal) model to be generated. The test data used to evaluate the impersonal models must *not* include data from any user also present in the training set.

In order to build and evaluate the personal models, data from each of the 51 subjects is separated, and then the data for each subject is partitioned into training and test sets using 10-fold cross validation. The results for personal models are based on the entire population of 51 users, and represent the performance averaged over the 51 users. The impersonal models are generated and evaluated very differently. In this case, the data from one user is separated and placed into the test set, while the data for the remaining 50 users is placed into the training set. A model is then built using the data from the panel of 50 users and is evaluated on the one "test" user; this is repeated 51 times so that all subjects are evaluated once. Based on this procedure, the impersonal models are generated from *much* more training data than the personal models—which is what we expect in realistic applications given the cost of generating personal training data.

Table 1 shows the activity recognition results for the personal models generated using the Random Forest classification algorithm. The accuracy of each of the eighteen activities is shown, for nine different sensor configurations. The first four configurations are for each of the individual sensors: the watch accelerometer, watch gyroscope, phone accelerometer, and phone gyroscope. Then multiple sensors are fused in an attempt to improve performance. These fused sensor configurations are: Watch (watch accelerometer and gyroscope), Phone (phone accelerometer and gyroscope),

Accels (phone and watch accelerometers), Gyros (phone and watch gyroscopes), and All (phone and watch accelerometers and phone and watch gyroscopes).

The results show that using all four sensors yields the best overall performance, although using the phone and watch accelerometers yields equivalent performance. Using these fused sensors does better than using any single sensor. Overall performance is quite good since when using all four sensors the average activity recognition performance, for the personal models, is 94.3%.

The results for impersonal models are presented in Table 2. The same nine sensor configurations are evaluated as with the personal models. As before, the best performance is achieved when using all four sensors, which yields an overall accuracy of 66.5%. The results for the impersonal models are much worse than for the personal models, even though the model is trained using much more data. While the overall activity recognition performance is quite low, certain activities, such as jogging, can still be recognized with relatively high accuracy.

The smartwatch sensors are particularly helpful for hand-based activities. To see this, consider the second grouping of activities, for both the personal and impersonal models, which begin with "Dribbling." For these seven activities, if we compare the accuracy results for the watch sensors against the results for the phone sensors, we see that in every case the watch sensors yield higher accuracy.

Based on these results, we can conclude that one can achieve highly accurate activity recognition results using only a smartphone and smartwatch, if personal models are built. The superiority of the personal models means that users move in different ways to perform the various activities, and that by exploiting these differences one can do much better at activity recognition. Personal models require the user to supply labeled training data, which entails some effort on their part, but this can be automated into a "self-training" phase, where the smartphone sequences the user through a set of activities. The results also show that the best results are achieved when the smartphone and smartwatch are both used. These results indicate that much more powerful activity tracking applications can be developed in the future, including some that might be better able to track eating activities.

4.4 Biometrics Experiments and Results

Biometrics can be used to identify or authenticate a person. In the context of this work, the identification task is to uniquely identify a user from a set of users using a sample of their motion sensor data. In contrast, the authentication task is simply to distinguish a user from an imposter. Identification is a multi-class learning problem while authentication is a binary class learning problem. Virtually all prior work on motion-based biometrics is based on gait—walking data is used as a biometric signature. In this study, each of the 18 different activities mentioned earlier are considered as biometric signatures. All experiments use stratified 10-fold cross-validation to build and evaluate the models (the stratification ensures that each fold contains the same distribution of users). Given that each example corresponds to

Table 1 Personal model activity recognition accuracy results

Activity	Watch		Phone		Watch	Phone	Accels	Gyros	All
	Accel	Gyro	Accel	Gyro					
Walking	87.8	85.6	95.8	92.3	89.1	96.6	96.8	94.4	97.0
Jogging	96.9	93.6	95.5	94.3	97.3	98.6	99.3	98.1	99.3
Stairs	85.5	70.4	89.9	84.1	84.0	92.7	93.7	88.3	93.8
Sitting	87.3	62.8	86.7	59.7	84.0	87.0	91.9	70.8	91.8
Standing	90.7	59.0	90.0	68.2	89.7	90.2	94.8	75.1	94.7
Kicking	82.9	72.7	87.8	80.4	84.4	90.6	93.3	86.1	92.7
Dribbling	91.2	90.6	84.9	75.7	96.1	88.2	95.2	94.7	95.6
Catch	90.5	88.7	83.2	73.9	94.4	85.8	94.3	94.2	94.7
Typing	94.1	83.3	90.3	69.2	92.9	92.3	95.8	83.3	95.4
Writing	89.9	77.6	89.7	67.6	91.2	90.8	92.4	81.2	92.9
Clapping	95.0	92.7	88.7	72.6	96.6	91.0	96.8	94.1	97.6
Teeth	91.9	81.6	90.0	69.6	94.8	90.4	96.2	86.1	95.2
Folding	89.8	85.3	88.4	82.9	95.2	92.1	95.9	95.6	96.1
Pasta	83.3	68.3	84.4	48.0	84.1	85.8	92.2	70.4	92.6
Soup	86.6	69.1	86.3	52.6	87.3	85.5	93.0	74.3	93.9
Sandwich	72.7	50.5	86.7	48.1	70.9	84.7	91.1	59.1	90.4
Chips	78.8	60.6	82.9	50.1	80.0	83.0	92.0	69.9	92.4
Drinking	80.9	65.2	85.5	50.0	80.8	85.2	92.7	69.9	92.1
Ave	87.5	75.4	88.2	68.8	88.5	89.5	94.3	82.5	94.3

Table 2 Impersonal model activity recognition accuracy results

Activity	Watch		Phone		Watch	Phone	Accels	Gyros	All
	Accel	Gyro	Accel	Gyro					
Walking	62.3	66.4	35.2	44.7	71.4	58.6	68.4	68.5	73.5
Jogging	92.2	76.3	68.8	64.6	91.6	84.4	97.6	82.3	96.4
Stairs	65.1	47.2	42.8	56.3	66	71.3	72.8	62.3	75.7
Sitting	54.6	50.3	21.2	19.6	58.8	23.4	58.4	52.4	58.1
Standing	73.3	48.4	38.3	37.9	69.7	61	79.2	57.9	80.2
Kicking	70.5	53.1	43.9	50.3	70.4	57.3	78.8	60.8	80.6
Dribbling	59.2	63.2	34.9	31.2	70.7	34.1	70.2	73.8	74.6
Catch	65.1	63.5	37.4	30.1	81.9	35.7	78	74	78.8
Typing	57.7	58.1	23.3	12.5	68.4	12.2	58.8	56.7	64.2
Writing	61.6	58	19.1	10.5	74.4	9.2	69.4	63.9	72.4
Clapping	65.9	66	13.2	13.8	72.6	14.3	78.1	69.3	75.3
Teeth	64.4	52.8	26.3	9.7	71	17.7	69.7	52.5	71.9
Folding	73.9	72.5	22.4	33.3	82.6	33.8	79.5	80	87.2
Pasta	44.1	43.1	17.9	10.1	53.1	10.4	44.1	42.3	48.2
Soup	45.8	33.4	13.9	13.2	53.9	7.5	47.9	33.1	48.2
Sandwich	17.4	10.9	13.4	4.5	17.8	8.2	15.1	11.7	14.3
Chips	39.7	31.9	14.1	10.6	46.1	12.4	40	32.8	44.7
Drinking	46.9	44.8	12.6	7.6	55	5.3	48.5	42.1	52.6
Ave	58.9	52.2	27.7	25.6	65.3	30.9	64.1	56.5	66.5

10 s of data, the most natural application of this work would yield results based on a single 10-s sample. For biometrics, we can assume that the sensor data that is collected from a device is from one person. Thus, we are free to use more than one example for identification or authentication purposes. In the results in this case study, each decision is based on five examples (50 s of data) and a majority voting scheme is used. The majority voting scheme yields significantly improved results.

Table 3 shows the identification results using the majority voting strategy. The random forest algorithm was used to generate the individual classifiers. As with the activity recognition case study, the results are reported for each of the nine sensor configurations—and as before the best results occur when using either all four sensors ("All") or the accelerometers on both the smartphone and smartwatch ("Accels"). The identification accuracies are very high and most activities yield good results. This includes walking, which is the standard activity used for motion-based biometrics. But even the eating activities yield good results, which indicates that people eat in very distinctive ways. Note that these identification results are based on a pool of 51 subjects, so that a strategy of guessing a person's identity would yield an accuracy just under 2%. The performance would undoubtedly degrade with a larger pool of subjects, so it would be interesting to extend this study to include a much larger pool of subjects.

An authentication model must distinguish a specific user from an imposter, which means that each subject must have their own authentication model. The training data must include data from the subject to be authenticated, combined with data from a panel of other subjects, where this panel of other subjects serves as a set of imposters. In real-world situations, we cannot assume that the imposters trying to fool the authentication system would have provided data for training, so in this scenario the test set is made up of data from subjects not represented in the training set. Given that there are 51 total subjects, the 50 "other" subjects are partitioned into two sets, one set to be used in the training set and the other set to be used in the test set. Since authentication is a binary classification problem and the positive class (the user to be authenticated) is rare in comparison to the negative class (imposters that should be rejected), the panel of other subjects was under-sampled to create a training set that is made up of only 75% imposters (several different proportions were evaluated but only had a minimal impact on the results).

Table 4 shows the authentication results. As with the prior identification results, the Random Forest algorithm was used to induce the models and majority voting is used to boost performance. Equal Error Rate (EER) is typically used to assess authentication performance and is used in Table 4. EER is calculated as the point where the False Acceptance Rate (FAR) equals the False Rejection Rate (FRR). FAR is the rate at which the model incorrectly accepts an imposter as a legitimate user, while FRR is the rate at which the model incorrectly rejects a legitimate user. The results in Table 4 again show that the best results are achieved when either all four sensors are used or both accelerometers are used. The results also show that walking is a very good activity for authentication purposes—when using all four sensors, walking yields an equal error rate of 6.8%, which is the second lowest EER (eating pasta appears to give better results but is clearly not the most practical activity).

Table 3 Identification accuracy performance per activity (with voting)

Activity	Watch		Phone		Watch	Phone	Accels	Gyros	All
	Accel	Gyro	Accel	Gyro					
Walking	94.1	80.4	100.0	100.0	90.2	100.0	100.0	100.0	100.0
Jogging	90.0	88.0	100.0	100.0	98.0	100.0	100.0	100.0	100.0
Stairs	70.0	43.8	98.0	90.0	75.0	96.0	100.0	91.7	100.0
Sitting	88.2	33.3	100.0	62.7	86.3	98.0	100.0	64.7	100.0
Standing	82.4	20.0	98.0	39.2	84.0	94.1	100.0	50.0	100.0
Kicking	76.0	32.0	96.1	68.6	82.0	100.0	100.0	80.0	98.0
Dribbling	98.0	90.2	96.1	68.6	86.3	98.0	96.1	96.1	100.0
Catch	78.0	85.7	98.0	70.0	91.8	100.0	100.0	91.8	100.0
Typing	94.0	50.0	100.0	89.8	100.0	100.0	100.0	95.9	100.0
Writing	94.1	58.8	96.1	80.0	98.0	100.0	100.0	90.0	100.0
Clapping	96.1	90.2	100.0	86.3	98.0	100.0	100.0	100.0	98.0
Teeth	94.1	62.7	98.0	82.4	96.1	100.0	100.0	94.1	100.0
Folding	64.7	39.2	100.0	76.5	86.3	96.1	100.0	78.4	100.0
Pasta	84.0	48.0	100.0	56.0	84.0	100.0	100.0	71.4	98.0
Soup	88.2	62.0	100.0	66.7	88.0	100.0	100.0	80.0	100.0
Sandwich	84.0	38.0	98.0	68.0	82.0	100.0	100.0	73.5	98.0
Chips	82.4	41.2	100.0	76.0	82.4	98.0	98.0	80.0	100.0
Drinking	86.3	41.2	100.0	58.8	80.4	100.0	100.0	60.8	100.0
Ave	85.8	55.8	98.8	74.4	88.3	98.9	99.7	83.2	99.6

An activity like clapping yields fairly good results, and since it could reasonably be performed in most environments, perhaps should be considered for authentication.

The biometric results, for both identification and authentication, have been presented in terms of individual activities. A long term goal is to build a system that performs continuous biometrics—that continuously validates a person's identity in the background, as the person goes about their normal tasks, without requiring the user to perform any specific activity. Thus, in the next set of identification experiments, we move a step closer to continuous biometrics by using the set of all 18 activities—without explicitly labeling each activity. Thus, the question becomes whether we can identify someone based on the sensor data generated from a diverse set of unlabeled activities. This is only a step towards continuous biometrics, since only eighteen activities are considered, rather than all activities a person might perform during their daily activities.

Three variations of the basic experiment are conducted. The first is the basic experiment: activity labels are not provided. The second experiment provides the activity labels and is provided for comparison purposes, to assess the impact of not having the activity labels. The third experiment does not include any activity labels, but uses an activity recognition model to predict the activity. Thus, this is a two stage approach, where the activity label is predicted and then is used in the biometric identification process. Table 5 provides the results for all three variations of the experiment, and also includes the results when the majority voting strategy is not used and is used, corresponding to the situation of making an identification using a 10 s sample of data or a 50 s sample of data, respectively.

The results in Table 5 demonstrate that good performance is possible even without the activity labels, at least when the voting strategy is used. Specifically, the results indicate that identification accuracy is 99.1% when all sensors are used and majority voting is used, even when no labels are provided. In fact, the results show that including the labels in this situation yields the same accuracy. Somewhat surprisingly, in this case predicting the activity label yields slightly worse performance than not having it. The key conclusion from Table 5 is that it is possible to achieve good identification accuracy when the input is only an unlabeled stream of activity data.

Based on the results in this section, motion-based biometrics using a smartphone and/or smartwatch can be effective. Wearable computing applications that perform much more granular activity recognition should arrive over the next decade, as should applications that use a person's motion to passively perform biometric identification.

5 Summary and Future Directions

This chapter provides a basic overview of wearable computing. It began by defining wearable computing and emphasized key characteristics, including that wearable computing technology should be easy to utilize without much conscious effort. It then provided a tour through the history of wearable computing devices, and in doing so demonstrated the diversity of wearable computing and wearable computing

Table 4 Authentication equal error rate per activity (with voting)

Activity	Watch		Phone		Watch	Phone	Accels	Gyros	All
	Accel	Gyro	Accel	Gyro					
Walking	13.2	17.2	9.4	9.8	13.9	8.8	11.3	10.0	6.8
Jogging	16.2	15.2	7.8	10.8	12.7	9.7	9.0	11.2	8.3
Stairs	19.3	23.9	13.4	12.5	18.9	9.3	8.4	14.1	6.9
Sitting	14.5	32.1	10.4	23.7	17.0	8.8	10.0	21.1	10.2
Standing	16.7	31.6	12.1	22.1	15.2	10.9	10.0	21.5	7.7
Kicking	21.0	24.1	10.6	19.4	16.6	11.0	10.1	18.8	11.0
Dribbling	16.4	16.1	10.3	21.0	14.5	9.7	10.0	11.8	11.5
Catch	16.3	15.5	9.7	19.3	14.9	10.0	9.3	13.9	10.0
Typing	13.0	20.7	8.3	15.4	14.0	8.9	8.6	13.3	8.8
Writing	10.7	21.3	8.7	15.7	11.6	9.2	9.0	16.0	10.1
Clapping	12.9	17.2	9.4	13.4	13.2	10.1	8.1	14.8	8.5
Teeth	13.3	20.0	10.1	14.0	14.4	10.2	10.8	14.9	8.2
Folding	17.0	23.4	7.9	18.6	17.3	10.0	8.1	16.2	7.1
Pasta	14.3	26.6	8.0	23.7	18.5	8.9	9.0	19.6	5.4
Soup	17.0	22.3	7.3	19.2	13.3	6.1	7.8	17.5	8.0
Sandwich	17.5	25.7	9.9	17.9	17.7	11.4	8.2	16.2	9.3
Chips	14.7	25.9	9.9	21.5	18.1	10.3	8.5	17.2	8.0
Drinking	16.6	25.1	11.3	19.2	13.9	10.2	10.9	19.9	8.1
Ave	15.6	22.4	9.7	17.6	15.3	9.6	9.3	16.0	9.3

Table 5 Identification accuracy using all eighteen activities

Sensors used	Without label		With label		Predicted label	
	voting?		voting?		voting?	
	No	Yes	No	Yes	No	Yes
Phone accel	58.0	96.8	58.5	97.6	30.3	96.0
Phone gyro	27.4	61.6	28.6	65.1	27.0	63.1
Watch accel	27.8	76.0	28.6	77.3	62.7	75.4
Watch gyro	12.4	39.8	13.2	43.9	51.8	42.4
Phone	61.2	97.0	62.1	97.5	32.7	96.2
Watch	28.6	77.1	29.3	77.9	66.6	80.6
Accel	64.0	99.2	63.9	99.3	64.0	98.9
Gyro	30.3	72.3	30.6	73.0	56.3	72.9
All	64.7	99.1	65.1	99.1	67.0	98.9
Ave	41.6	79.9	42.2	81.2	43.8	80.5

applications. The history also demonstrated that many of the more recent wearables have their roots in wearables that were developed decades ago. Lessons learned from the history of wearables were presented and used to predict future trends in wearable computing. The chapter then described several wearable common application areas, and the industries that are currently benefitting most from this technology. Case studies on activity recognition and biometrics were provided to demonstrate how some wearable computing applications are implemented, and highlight how data science methods can lead to more powerful future wearable computing applications.

Wearable computing has entered the mainstream over the last few years, first with the introduction of activity trackers and then with smartwatches. These devices have only begun to tap the potential of wearable computing and even they have not yet completely proven themselves. For example, activity trackers are quite popular, but have not been around quite long enough to prove that they are not a fad—and there are signs that even the commercial success of Fitbit may be fading. In fact, there are even studies that indicate that the benefits of fitness tracking may be overblown. One study showed that adding fitness tracking to a standard behavioral intervention for weight loss resulted in a *reduction* in weight loss [35]. Similarly, many users find the benefits of using a smartwatch to be minimal, and the growth of the smartwatch market has thus far been rather disappointing. More ambitious wearable computing devices, like Google Glass, still face an uncertain future—Google Glass itself was withdrawn from the commercial market until the product can be improved. It is still unclear whether augmented reality devices will ever enter the mainstream.

However, there is reason for optimism. There is still great interest in wearable computing and devices like Google Glass may simply have been released prematurely. After all, the Apple Newton PDA was a flop, but ultimately Apple released the iPhone, which includes much of the PDA functionality, and it was a tremendous success. Furthermore as the electronics continue to shrink and power requirements

are reduced, wearable computing devices will become cheaper and less cumbersome. This is especially true for wearable sensors, which could ultimately be embedded on our clothing. Medical applications of wearable computing could alone turn out to have enormous benefits. Thus, there is great potential for growth in wearable computing technology, but such technology must provide substantial and concrete benefits to the user.

References

1. DARPA: Proceedings of the Wearables in 2005 Workshop (1996)
2. Mann, S.: An historical account of the 'WearComp' and 'WearCam' inventions developed for applications in 'personal imaging'. In: First International Symposium on Wearable Computers, 13–14 Oct, 1997. IEEE. Cambridge, MA, USA (1997)
3. Hayes, J.: XII things you (Probably) didn't know. Eng. Technol. **8**(12), 39–42 (2013)
4. Thorpe, E.O.: The invention of the first wearable computer. In: Proceedings of the Second International Symposium on Wearable Computers (SWC), pp. 4–8 (1998)
5. Bass, T.A.: The Eudaemonic Pie. Houghton Mifflin (1985)
6. Casio: Casio History 1980. Web (2017). http://world.casio.com/corporate/history/chapter02/. Accessed 13 June 2017
7. Sutherland, I.: A head-mounted three dimensional display. In: Proceedings of the Fall Joint Computer Conference, pp. 757–764. IEEE CS Press, Los Alamitos, California (1968)
8. Mann, S.: Wearable computing: a first step toward personal imaging. Computer **30**(2) (1997)
9. Starner, T.: Project glass: an extension of the self. IEEE Pervasive Comput. **12**(2), 14–16 (2013)
10. Rhodes, B., Starner, T.: Remembrance agent: a continuously running automated information retrieval system. In: The Proceedings of the First International Conference on the Practical Application of Intelligent Agents and Multi Agent Technology (PAAM), pp. 487–495 (1996)
11. IEEE Computer Society: Digest of Papers from the First International Symposium on Wearable Computers, 13–14 Oct (1997). IEEE Computer Society (1997)
12. Edwards, F.: Introducing the Tiny, Wearable 12MP Perfect Memory Camera. DigitalRev, Web (2016)
13. D'Cruze-Sharpe, R.: Best cameras for lifelogging: how to livestream your life on the move. In: Wearable (2016). www.wareable.com/cameras/best-wearable-lifelogging-cameras. Accessed 3 July 2017
14. Kwapisz, J.R., Weiss, G.M., Moore, S.A.: Activity recognition using cell phone accelerometers. ACM SIGKDD Explor. **12**(2), 74–82 (2010)
15. Weiss, G.M., Timko, J.L., Gallagher, C.M., Yoneda, K., Schreiber, A.J.: Smartwatch-based activity recognition: a machine learning approach. In: Proceedings of the 2016 IEEE International Conference on Biomedical and Health Informatics, Las Vegas, NV, pp. 426–429 (2016)
16. Streitfeld, D.: Google Glass picks up early signal: keep out. The New York Times. Retrieved June 26, 2013
17. Jhajharia, S., Pal, S.K., Verma, S.: Wearable computing and its application. Int. J. Comput. Sci. Inf. Technol. **5**(4), 5700–5704 (2014)
18. Azuma, R.T.: A survey of augmented reality. Presence: Teleoperators Virtual Environ. **6**(4), 355–385 (1997)
19. Brooks, F.P.: The computer scientist as toolsmith II. CACM **39**(3), 61–68 (1996)
20. Stewart-Smith, H.: Education with augmented reality: AR textbooks released in Japan, ZDNet Apr. 4, 2012. http://www.zdnet.com/article/education-with-augmented-reality-ar-textbooks-released-in-japan-video
21. Blum, T., Kleeberger, V., Bichlmeier, C., Navab, N.: Mirracle: an augmented reality magic mirror system for anatomy education. In: Proceedings of the 2012 IEEE Virtual Reality, pp. 115–116. IEEE Computer Society (2012)

22. Kay, M., Santos, J.: Takane M (2011) mHealth: new horizons for health through mobile technologies. World Health Organ. **64**(7), 66–71 (2011)
23. MiniMed 530G System: Medtronic, July 1, 2017. http://www.medtronicdiabetes.com/products/minimed-530 g-diabetes-system-with-enlite
24. Muensterer, O.J., Lacher, M., Zoeller, C., Bronstein, M., Kübler, J.: Google Glass in pediatric surgery: an exploratory study. Int. J. Surg. **4**, 281–289 (2014)
25. Kaddoum, R., et al.: A randomized controlled trial comparing the AccuVein AV300 device to standard insertion technique for intravenous cannulation of anesthetized children. Pediatr. Anesth. **22**(9), 884–889 (2012)
26. Miyake, R.K., Zeman, H.D., Duarte, F.H., Kikuchi, R., Ramacciotti, E., Lovhoiden, G., Vrancken, C.: Vein imaging: a new method of near infrared imaging, where a processed image is projected onto the skin for the enhancement of vein treatment. Dermatol. Surg. **32**, 1031–1038 (2006)
27. Shu, C.: Startup brain power uses Google Glass to develop apps for kids with autism. TechCrunch Dec. 23, 2014. Accessed June 30 2017
28. Metz, R.: Augmented-reality glasses could help legally blind navigate. MIT Technology Review, June 15, 2015. www.technologyreview.com/s/538491/augmented-reality-glasses-could-help-legally-blind-navigate
29. Conrad, C., Fusaglia, M., Peterhans, M., Lu, H., Weber, S., Gayet, B.: Augmented reality navigation surgery facilitates laparoscopic rescue of failed portal vein embolization. J. Am. Coll. Surg. **223**(4), e31–e34 (2016)
30. Lockhart, J.W., Pulickal, T., Weiss, G.M.: Applications of mobile activity recognition. In: Proceedings of the ACM UbiComp International Workshop on Situation, Activity, and Goal Awareness, Pittsburgh, PA (2012)
31. Bao, L., Intille, S.: Activity recognition from user-annotated acceleration data. Lecture Notes Computer Science 3001, pp. 1–17 (2004)
32. Google Glass: Wearable Tech in the Warehouse: Harvard Business School, Nov 7, 2017 (2017). https://rctom.hbs.org/submission/google-glass-wearable-tech-in-the-warehouse/
33. Gans, E., Roberts, D., Bennett, M., Towles, H., Menozzi, A., Cook, J., Sherrill, T.: Augmented reality technology for day/night situational awareness for the dismounted soldier. In: Proceedings of the SPIE 9470, Display Technologies and Applications for Defense, Security, and Avionics IX; and Head- and Helmet-Mounted Displays (2015)
34. Yoneda, K., Weiss, G.M.: Mobile sensor-based biometrics using common daily activities. In: Proceedings of the 8th IEEE Annual Ubiquitous Computing, Electronics & Mobile Communication Conference, New York, NY, pp. 584–590 (2017)
35. Jakicic, J.M., Davis, K.K., Rogers, R.J., King, W.C., Marcus, M.D., Helsel, D., Rickman, A.D., Wahed, A.S., Belle, S.H.: Effect of wearable technology combined with a lifestyle intervention on long-term weight loss: the IDEA randomized clinical trial. JAMA **316**(11), 1161–1171 (2016)

Wearables Security and Privacy

Jorge Blasco, Thomas M. Chen, Harsh Kupwade Patil and Daniel Wolff

Abstract Wearable devices equipped with various embedded sensors are finding many applications in health care and other sectors. As a relatively new class of mobile computing, there is little experience with security and privacy problems. This chapter aims to bring attention to these important but somewhat overlooked issues. We describe the components in wearables (sensors, processors, software, and communications) and highlight the security issues related to wireless protocols, vulnerabilities, and privacy.

1 Introduction

In the past decade, smartphones have become a ubiquitous platform for mobile computing, allowing users to carry around serious computing power and always-on Internet connectivity [31]. Wearable devices extend mobile computing to be worn on the body which offers some appealing advantages: they can be carried around conveniently and continuously; they can be operated mostly hands-free; they can be highly

J. Blasco
Royal Holloway, University of London, London, UK
e-mail: jorge.blascoalis@rhul.ac.uk

T. M. Chen (✉) · D. Wolff
City, University of London, London, UK
e-mail: tom.chen.1@city.ac.uk

D. Wolff
e-mail: Daniel.Wolff.2@city.ac.uk

H. Kupwade Patil
San Jose Laboratory, LG Electronics, Santa Clara, CA, USA
e-mail: harsh.patil@lge.com

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_11

personalized in a variety of form factors; and they can incorporate an array of sensors to measure health signs [32, 45] and personal activities [16, 36, 38].

Wearables are becoming increasingly popular in sectors including infotainment, fitness, health care, and industry [15]. Statistica [63] estimates that 85 million wearables were shipped in 2015, which will increase by 58% to 135 million in 2016, and then to 190 million in 2017. Gartner [24] predicts that 50 million smartwatches, 35 million wristbands, 24 million sports watches, and 21 million other fitness monitors will be sold worldwide in 2016. The numbers do not include wearable systems specialized for military applications [71].

Wearables for infotainment include smart glasses, heads-up displays, and smartwatches. Fitness and healthcare applications involve wristbands, smart garments, chest straps, and sports watches. Wearables for industry and military applications include head-mounted displays and hand-worn terminals. Other forms of wearable devices are gloves, shoes, contact lenses, armbands, rings, caps, bracelets, and earbuds. Wearables are often designed with multiple functions, e.g., smartwatches and wristbands can monitor fitness, make contactless payments, receive or send messages, wirelessly unlock doors, and perform many more things depending on software apps.

Although wearables have certain advantages over smartphones, wearables are more likely to complement smartphones than replace them. Wearables extend computing to the body but are constrained by their often small size and mobility requirements [19]. They typically must be designed to minimize battery power usage [61]. Their hardware resources are limited usually in terms of memory and computing. Their wireless communication range is short mainly to save energy. For these reasons, they often work with smartphones to take advantage of the phone's greater computing and communications capabilities. An example of a healthcare scenario is shown in Fig. 1, but this is not a unique configuration. In this example, a smartphone may act as a hub to collect and process data from wearable sensors [42, 69]. Hubs have relatively large data storage, powerful processors, and broadband Internet connectivity. Hubs may carry out lightweight signal processing on the data and transmit

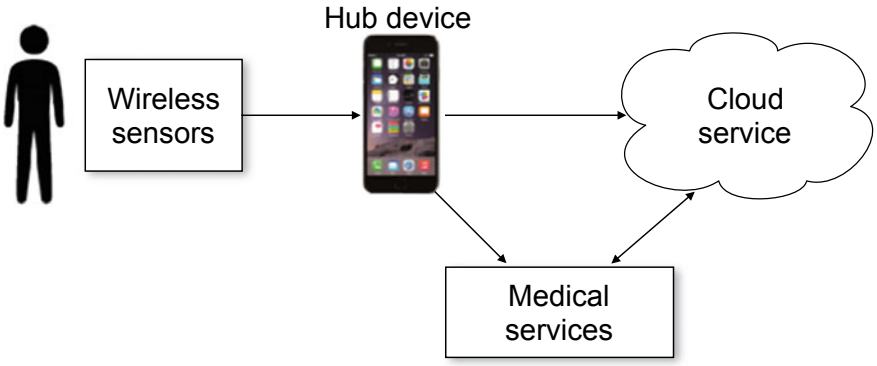


Fig. 1 A healthcare scenario

a fraction of the data to cloud servers for powerful analysis and long-term storage. Data may be shared with authorized medical services (e.g., doctors, hospitals, etc.). In the long term, wearables will be more standalone devices, as suggested by recently introduced smartwatches with Long-Term Evolution (LTE) cellular capabilities.

In the bigger picture, wearables (and smartphones) will be a part of the expanding Internet of Things (IoT) [51]. The IoT will be made up of a massive number of interconnected "smart" objects with sensing, communications, and information processing capabilities [47]. However, IoT solutions are being designed with security as a secondary consideration [53]. Security and privacy concerns for the IoT are relevant to wearables as well [67]. For example, personal health data collected by a wearable might be stolen for malicious purposes, or a vulnerability in a wearable device might be exploited by ransomware to force the owner to pay a ransom. Some wearables are used as authentication devices (e.g., for payments) which make them attractive targets for criminals.

Although research in wearables has been ongoing for decades, they have become mainstream popular with consumers only in recent years [52, 62]. There is little experience with security issues at the current time. Wearables increase the risk of certain security and privacy issues because of the following reasons:

- Wearables have a variety of biosensors, which can collect a great amount of personal data about a person [59];
- Wearables are worn constantly so a person may be monitored continuously;
- Wearables are always network-connected and accessible;
- Wearable devices are often designed for functionality and price instead of security.

In a real sense, wearables are the most intimate "personal" computing devices because they know a person's activities and physiology. It is easy to see that wearables will be attractive targets for criminals, not only for the valuable personal data stored in them but also for other possible attacks:

- Attack scenario 1: Wearables are used for access control (to open locks or log into computer accounts). A wearable is identified by a unique cryptographic key stored in memory. A criminal steals a wearable to gain entry to a victim's house or bank accounts.
- Attack scenario 2: A criminal gains access to the sensor data in a wearable to steal a victim's biometric data, e.g., facial image, voice pattern, and heart rate data. Using the stolen biometric data, the criminal carries out identity theft by masquerading as the victim.
- Attack scenario 3: A criminal eavesdrops on wireless transmissions from the wearable to steal personal data.
- Attack scenario 4: A criminal takes control over the wearable device (e.g., locks the wearable) and extorts the victim for money in return for giving back control.
- Attack scenario 5: A criminal takes control over the wearable device, perhaps with malware, and uses its resources for malicious purposes, e.g., spam, botnet, or a stepping stone to launch attacks on other devices.

The aim of this chapter is to bring more attention to security and privacy issues for wearable devices. Section 2 begins with a description of wearable devices and their components. Section 3 examines the security of common wireless protocols that are being implemented in wearables. Section 4 describes the vulnerabilities of wearable devices. Finally, Sect. 5 reports on privacy issues.

2 Wearable Devices

What is a wearable device? Wearables are a broad class of mobile computing devices with significant power and size limitations imposed by the form factors. It may be easiest to think of traditional wearable objects—such as clothes, watches, rings, glasses, and headgear—and add computing and communications capabilities to make a wearable device. Thus, in contrast, smartphones are not in the class of wearable devices because phones are traditionally thought to be "carried" but not "worn." Like any computer, wearables have processors, memory, and software. They may or may not be connected to the Internet, depending on their application. Since they are worn continuously and close to the body, they tend to include an interesting array of sensors for monitoring a range of biosignals [59]. Valuable physiological data can be collected over long time frames that can be analyzed for baseline patterns, anomalies, and gradual progression of certain symptoms.

In this section, we describe four major components in wearable devices: sensors, signal processing, processors, and software. While this section is intended mostly for background, security risks and vulnerabilities are pointed out where appropriate.

2.1 Sensors

A wide variety of sensors can be accommodated in wearable devices [10, 42]. The cost-effective production of small sensors is now possible due to technological advances in microelectronics, materials, optics, and miniaturization. Typical wearable sensors are noninvasive, i.e., work outside of the human body, and directly on the skin or in very close proximity. Invasive sensors are preferable for measurements of internal processes (e.g., bile sensors [9]) but involve surgical implantation or ingestion which are naturally unappealing.

The description of sensors here aims to be comprehensive for two reasons. First, the variety of sensors embedded in wearables is one of the major differences between wearables and traditional computers (including smartphones). Second, the data collected from sensors poses new security risks such as loss of privacy of very personal data (related to physiology, medical conditions, and daily activities) and valuable biometric data that might be stolen for purposes of identity theft.

2.1.1 Light Sensors

Cameras: Digital cameras are optical sensors for taking images or videos, combined with other sensors, special circuitry, and sophisticated signal processing for enhancing the picture quality (e.g., to compensate for low light, shaking, and motion, as well as recognize faces). They are commonplace now in smartphones, smart glasses, and other wearables. A wide range of applications include infotainment, augmented reality, and biometrics (face, retina, and fingerprint recognition).

Cameras are used in older types of fingerprint scanners to capture an image, and then the algorithms analyze the light and dark areas to recognize patterns such as ridges. An array of LEDs provides lighting for the fingerprint at scan time. This type of optical fingerprint scanner has been shown to be vulnerable to spoofing by high-quality images of stolen fingerprints. More modern fingerprint scanners are capacitive which are more difficult to fool.

Face recognition technology has been around for several decades, and many techniques are available, e.g., Viola–Jones algorithm, principal component analysis, independent component analysis, linear discriminant analysis, and so on. Face recognition is not as popular for smartphones as fingerprint recognition perhaps because face recognition is generally less reliable (affected by shadows, occlusions, and so on) and easier to spoof in the sense that faces are easier to steal than fingerprints.

For biometrics, iris patterns (the colored ring in the eyeball between the central pupil and the sclera) are appealing because they do not change after age two. While the color of the iris is determined by genetics, the patterns in the ligaments of the iris are created by random tissue folding during gestation and are unique to each eyeball. Also, there are 225 different points of comparison that are unique to each iris, compared to 40 in a fingerprint. In general, a near-infrared (NIR) light is shown into the eye because it does not cause discomfort, unlike visible light. A separate camera is used to capture the image because standard digital cameras include infrared-blocking filters. Alternatively, some iris recognition systems look at the pattern of blood vessels in the white part of the eye.

Cameras offer a noncontact approach to measuring respiratory rate, in contrast to contact approaches requiring sensors on the chest and abdomen to measure movements there. Generally, cameras capture a video of a person in visible or infrared light, and the frames are analyzed to pick out the rhythmic movements indicative of exhalation and inhalation.

As a potential point of attack, cameras are an attractive target for criminals. Gaining access to the camera can allow theft of highly personal images and biometric data.

PPG: The photoplethysmograph (PPG) measures the pulse wave as the volume change of blood [64]. It takes advantage of the fact that blood absorbs infrared light. Typically, light is emitted by one or multiple LEDs on the skin; a photodetector on the same side will detect the scattered light or a photodetector on the other side will detect the transmitted light. Each time the heart beats, a blood pressure pulse is generated and propagated in the blood vessel. A local increase of blood pressure causes an increase in light absorption and attenuation of the light transmitted through

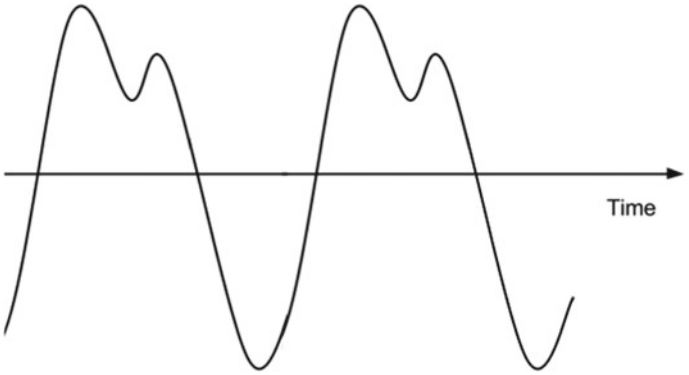


Fig. 2 An example of PPG signal

the tissue or reflected. An example of a PPG signal is shown in Fig. 2. Common operating wavelengths are between 510 mm (green) and 920 mm (infrared). Green works best on light skin and normal temperatures, whereas longer wavelengths are better for dark skins or cold temperatures. PPG is useful for monitoring heart rate [6], blood oxygen saturation (SpO_2), blood pressure, and stroke volume [54].

PPG sensor data may not be that valuable to criminals as health data, but heart rate is starting to be used for biometric authentication. PPG data may therefore be targeted for identity theft.

Pulse oximeter: A pulse oximeter is a device usually on the fingertip or earlobe (for their small capillaries) that works in a similar way as PPG. Two wavelengths of light are shown through the finger or earlobe to a photodetector on the other side to measure the fraction of oxygen saturation level in blood. The two wavelengths measure the absorption coefficients due to the difference in concentration of hemoglobin and deoxyhemoglobin levels in blood.

Blood pressure: The traditional method of measuring blood pressure is the sphygmomanometer, an inflatable cuff that squeezes the upper arm. Wearables offer a challenge to measure blood pressure with a much smaller apparatus. One approach is a cuff around the finger that applies a varying pressure. At the same time, infrared light is shown through the finger to a photodiode. Since the wavelength is primarily absorbed by hemoglobin, the light intensity fluctuations give information about the area of the finger cross section occupied by blood. The volume of the blood is related to pressure, so the light intensity can be related to arterial blood pressure.

Blood glucose: Light is one of the means to measure blood glucose concentration (other methods are described later). Diabetes has no immediate cure, and thousands of people are diagnosed each day. There are many options for monitoring glucose levels [66]. Light sensors that fit within a wearable device offer a noninvasive way that is clearly preferable to traditional invasive and painful ways.

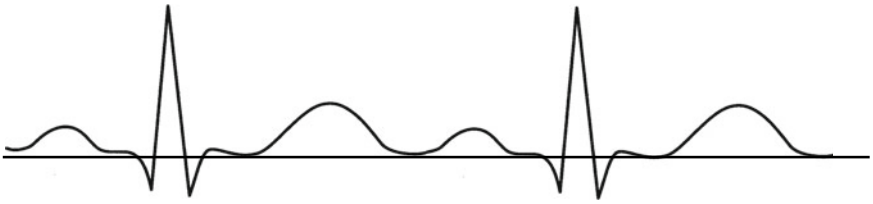


Fig. 3 An example of ECG signal

2.1.2 Electrical Sensors

ECG: Many wearable sensors focus on monitoring the cardiovascular system because of the electrical activity of the heart [59]. The most familiar electrical sensor is the electrocardiogram (ECG) consisting of two or more metal electrodes that must be in direct contact with the skin, usually facilitated by a gel for a proper connection [14]. They can be placed across the chest, wrists, and ankles. ECG electrodes measure the tiny voltage changes on the skin that arise from the pattern of depolarizing and repolarizing during each heartbeat. A healthy heart has a regular progression of depolarization starting with pacemaker cells in the sinoatrial node and eventually ending in the ventricles that create a typical ECG wave as shown in Fig. 3. Repolarization is a phase when cells return to a resting negative charge.

An ECG provides a large amount of information about the structure and function of the heart. Aside from a check of general health, it is useful for diagnosis of breathing difficulties, heart problems, fainting, seizures, and emergency situations. Wearables allow continuous ECG monitoring, which is particularly useful for people who are critically ill, undergoing general anesthesia, or have an infrequently occurring abnormal cardiac rhythm.

Unfortunately, many sources of noise can corrupt ECG signals: power line interference, electrode contact noise, motion artifacts, muscle contraction (refer to electromyogram below), and electromagnetic interference from other electronic devices. Practically, it is necessary to filter out all these noise sources.

Theft of ECG data may pose a serious privacy loss because ECG can reveal a substantial amount of information about a person's health and medical condition. Also, like PPG sensor data, heart rate measured by ECG (more accurate than PPG) may be valuable for biometric authentication. ECG data should be protected against identity theft.

Respiratory rate: Respiratory rate may be derived from an ECG because it has been observed that the respiration has a modulating effect on the ECG. The technique is called ECG-derived respiration (EDR) [41].

EMG: A surface electromyogram (EMG) is performed in a similar way as ECG with multiple electrodes on the skin to measure the electric potential generated by muscle cells when these cells are electrically or neurologically activated. A surface EMG is noninvasive but provides only limited information about muscle activity.

An intramuscular EMG gives a much more informative measurement but requires insertion of electrodes through the skin into the muscle tissue.

EEG: An electroencephalogram (EEG) measures voltage fluctuations resulting from ionic current within the neurons of the brain. Typically, multiple EEG electrodes are placed in a head-worn apparatus to make contact with the scalp. Noninvasive EEG is used to diagnose epilepsy, sleep disorders, coma, stroke, encephalopathies, and brain disorders in general. However, a clinical EEG can take 20–30 min; EEG is not good at measuring neural activity below the upper layers of the brain (the cortex), and generally the signal-to-noise ratio is poor.

Like ECG data, theft of EEG data may pose a serious loss of privacy. Unlike ECG data, the EEG is not currently used for biometric authentication, so the reason for theft of EEG data is not likely to be identity theft.

GSR: Another electrical sensor is the galvanic skin response (GSR or skin conductance) sensor used to measure the electrical conductance of the skin [44]. Two electrodes are placed on the skin close to each other and pass an imperceptible current between them. The measured electrical resistance of the skin depends on the moisture or sweat produced by the skin. Sweating is controlled by the sympathetic nervous system, and GSR is sometimes interpreted as an indicator of arousal or stress.

Temperature: Finally, electrical sensors are common for measuring temperature (among other methods such as infrared detection). Electrical temperature sensors can be built using a thermistor or thermocouple. A thermistor changes resistance with temperature; the resistance is measured by a bridge circuit containing the thermistor. A thermocouple takes advantage of the property that a small voltage is generated at a junction of different conductors that is proportional to their temperature difference.

2.1.3 Electrochemical Sensors

Sweat rate: A real-time sweat rate sensor was constructed from two capacitive humidity sensors at different distances from the skin [57]. A capacitive humidity sensor consists of a nonconductive foil which is covered with gold on both sides. The dielectric constant of the foil changes as a function of the relative humidity of the ambient atmosphere, which is measured as the capacitance value. The difference between the measurements at the two humidity sensors gives an indication of water vapor flow from the skin's surface.

Sweat: As mentioned earlier, sweat contains an abundance of interesting electrolytes and metabolites. Up to now, noninvasive biosensors have been able to monitor a single analyte at a time or lack on-site signal processing circuitry. Gao et al. [23] have built a wearable containing an array of electrochemical sensors for in situ sweat analysis including glucose, lactate, sodium, and potassium ions. The glucose and lactate sensors are electrodes coated with a specific enzyme, namely, glucose oxidase, and lactate oxidase, respectively. These enzymatic sensors generate electric current proportional to the abundance of the corresponding metabolites between the working electrode and a reference electrode.

Glucose: A noninvasive method to measure the glucose level in blood would be valuable for managing diabetes [66]. Optical methods to measure blood glucose were mentioned earlier. A correlation has been found between sweat glucose and blood glucose, so some researchers have focused on sweat glucose [43]. Sweat glucose may be measured noninvasively (as described above) but measurements can be easily confounded by other factors in sweat.

2.1.4 Motion Sensors

GPS: In wearables, motion sensors can be built based on location sensors or force-based sensors. GPS is a well-known satellite system for triangulating location on Earth using signals from four line-of-sight GPS satellites [13]. GPS receivers provide a location within a few meters or so, depending on the type of GPS receiver. Exposure of location information is sometimes seen as a threat to privacy.

Magnetometers: Magnetometers or compasses measure the direction of the Earth's magnetic field to determine the bearing or direction of an object. Digital magnetometers are small and inexpensive, and thus suitable for embedding in almost any electronic device including wearables. A digital magnetometer is a type of force-based motion sensor that is generally embedded within other force-based sensors such as accelerometers and gyroscopes.

Accelerometers: Accelerometers are widely used in smartphones and other mobile devices to detect device orientation and serve as input to motion-based games. Commonly used accelerometers measure g-force (1 g is 9.81 m/s^2) in the three axes: x, y, and z. Four kinds of accelerometers are available: piezoelectric, piezoresistive, capacitive, and servo-type sensors. They work on the principle of generation of electricity, change in resistance, change in capacitive effect, and change in heat induction, respectively.

Accelerometers along with gyroscopes may be used to infer a person's activities. Hence, the data may be considered to be worth protecting as personal data.

Gyroscopes: Gyroscopes measure attitude and rotation. Attitude is the orientation of the gyroscope relative to a point in space. By measuring changes in attitude, gyroscopes can also measure its rotation rate.

Pedometers: A pedometer counts the number of steps walked by detecting when a body tilts from side to side, e.g., by movement of the hips, and multiplies the number of steps by the length of each step to determine a total distance traveled. Inside a pedometer, a metal pendulum swings when the body tilts to one side to make electrical contact with an electronic counting circuit, incrementing the count by one. When the body tilts back, the pendulum swings back and breaks the circuit. Other pedometers are entirely electronic, using two or three accelerometers. These are arranged at right angles that detect minute changes in force when legs move during a step.

Shoe sensors: Shoes can be fitted with pressure sensors and accelerometers to track steps or analyze gait. Pressure sensors are usually made of several thin layers of a piezoresistive material, such as silicon, that becomes more resistant to an electric

current when force is put on it. The surface is connected to a Wheatstone bridge, which is designed to detect small differences in resistance.

2.1.5 Sound Sensors

Microphones: Microphones change sound waves into an electrical signal. They are inexpensive and small, so they are commonplace in many types of electronic devices. Microphones are useful for a variety of applications including voice recognition, respiration rate analysis, and emotion detection. Microphones have the drawback of capturing ambient noise as well as the interesting sound. Multiple microphones and signal processing techniques are typically used to reduce the effects of noise [11].

Microphones are often a target for criminals because access to sound may allow criminals to hear personal data or steal voice patterns for biometrics. Thus, the threats are privacy loss and identity theft.

Ultrasound: Ultrasound at frequencies above human hearing has many useful applications, e.g., fingerprint scanning. A fingerprint can be scanned by transmitting an ultrasonic pulse against the finger placed on a scanner. While some of the pulses are absorbed, the rest is bounced back to the sensor, depending on the ridges, pores, and other microstructures that are unique to each fingerprint. The sensor calculates the intensity of the returning ultrasonic pulse at different points on the scanner.

Fingerprint data should obviously be protected against theft by criminals who could use the data for identity theft.

2.2 Signal Processing

A wearable device will often send its sensor data to a hub device for long-term storage and heavy processing (e.g., data mining or classification). This saves memory storage and reduces energy consumption in the wearable. However, there are a number of functions that need to be carried out in the wearable, namely, signal conditioning and signal processing as shown in Fig. 4.

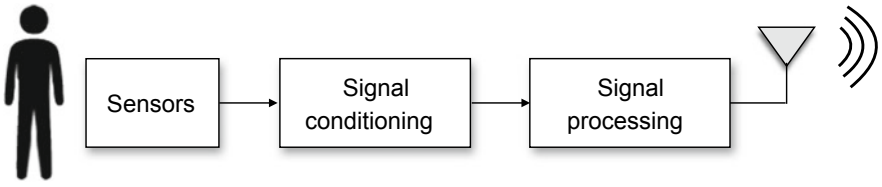


Fig. 4 Signal conditioning and signal processing in a wearable

Decisions about which functions to design into the wearable (as opposed to leave to the hub device or cloud service receiving the data from the wearable) are complicated by the following considerations:

- The total power consumed depends on the energy used by the sensors, sampling, signal preprocessing, and wireless transmission [59]. There are trade-offs, e.g., it might be more efficient sometimes to process data in the wearable instead of transmitting the data.
- Power consumption should be minimized but balanced against performance and cost, which depend on the application, for example, some applications may require a minimum sampling rate.
- Certain time-critical functions may be better to perform in the wearable, e.g., detection of imminent hazards. Another reason to carry out functions in the wearable might be unreliability of the wireless channel or cloud service.
- Wearables have a wide range of processing capabilities (from simple fitness bands to sophisticated smartwatches). Some functions may not be feasible to support in basic wearables.

Signal conditioning may include noise filtering or cancellation, signal amplification, anti-aliasing (e.g., low-pass filtering), and analog-to-digital conversion consisting of sampling and quantization [59]. Noise can be a substantial factor due to user movements, environmental noise, and changes in sensor locations (e.g., a smartwatch slipping around the wrist). Sampling frequency is another important consideration because a higher sampling rate not only improves data resolution but also increases the amount of data (and hence power consumption).

Signal processing may include data compression and lightweight classification but it is dependent on the application. Data compression reduces the total amount of data for transmission and storage, but lossy data compression achieves higher compression at the cost of discarding information that will be unrecoverable later. The compression algorithm depends on the application and what type of information should be discarded preferentially. In most cases, data should be transmitted to a hub device or cloud service for data mining and classification, but certain applications may necessitate lightweight classification to be performed in the wearable. For example, there may be applications that are sensitive to the communication delay or unreliability. In that case, algorithms for feature extraction and classification must be designed to be as efficient as possible [59].

2.3 Processors

Wearables take many forms depending on where they are worn on the body, but they are all constrained by power, memory, and computing resources. Understandably, people do not want to wear bulky, heavy equipment. At the same time, they have realistic expectations that wearables have limited functions. Hence, the

microprocessors found in wearable devices have lower specifications than desktop computers and laptops, and even some smartphones.

The most common processors are based on reduced instruction set computer (RISC) in contrast to traditional complex instruction set computer (CISC) processors designed for desktop computers, exemplified by Intel's x86 platform. The RISC approach chooses a set of simpler instructions than CISC in order to reduce the number of processor cycles required to perform each instruction, resulting in smaller hardware and lower power consumption. Some high-end wearable devices have a separate coprocessor to off-load the processing of sensor data from the main processor. A coprocessor referred to sometimes as a "sensor hub" is useful when the device has a great amount of sensor data that needs to be analyzed together in real time, requiring constant CPU attention.

While there have been many RISC processors (e.g., MIPS, SPARC, and PowerPC), processors based on the ARM architecture licensed from ARM Holdings have become the most popular, adopted in wearables as well as iOS and Android smartphones and tablets [39]. The ARM Cortex-M family is well suited for low-end wearables due to its small form factor and low power requirements. For instance, the Cortex-M3 processor is used in the Pebble watch, Fitbit fitness bands, and Arduino Flora. The Cortex-M4, Cortex-M7, and Cortex-M33 processors integrate digital signal processing (DSP) and floating point operations, which are advantageous for applications such as sensor fusion and power management.

The ARM Cortex-A processor family tends to focus graphics and CPU power, compared to the Cortex-M. This tends to be found in high-end wearables such as smartwatches supporting an operating system capable of running a variety of apps and communicating with other devices (like wireless earphones or smartphones).

Many wearables use custom systems on chip (SoCs) that usually integrate multiple cores, graphics processing unit (GPU), DSP, GPS, wireless communications, and support for audiovisual sensors. Well-known examples of SoCs include the following:

- Apple's 64-bit A9 (based on ARMv8) with an integrated M9 motion coprocessor that was first implemented in the iPhone 6S and 6S Plus;
- Samsung's multi-core Exynos 9 (also based on ARMv8) appearing in the Galaxy S8 and S8+ smartphones [58];
- Qualcomm's Snapdragon Wear 2100 based on ARM Cortex-A7 [49].

The ARM Cortex-M23, Cortex-M33, and Cortex-A series processors support TrustZone technology, which dedicates a secure area on the chip called trusted execution environment (TEE) [40]. The TEE is an area for trusted resources—software, data, and hardware—separated by hardware from untrusted resources. The trusted environment can also include memory, peripherals, interrupts, and bus transactions. Common uses include the protection of authentication mechanisms, cryptography, trusted software (e.g., secure boot and electronic wallet), and biometric data. Untrusted software cannot access secure resources directly. Thus, secure resources are protected from software attacks and common hardware attacks. Context switching between secure and nonsecure environments is done in software via a secure monitor call in Cortex-A, or hardware in Cortex-M.

Apple's A7 and later SoCs are based on ARMv8 and contain a secure coprocessor called "secure enclave" that is likely using ARM's TrustZone technology. The secure enclave is known to protect data from the Touch ID fingerprint sensor. Reportedly, it has its own secure boot process to ensure security. It has a unique, unalterable ID useful for creating a temporary encryption key to encipher its memory. It also contains an anti-replay counter.

2.4 Software

Wearable devices are highly fragmented from a software perspective, with many operating systems (OSs) sharing the market [3, 33]. An OS for wearables is different from a traditional desktop OS in a number of ways: it should optimize process scheduling and power consumption; it should support the wearable's user interface; it should optimize graphics processing; and it should support the wearable's sensors input/output. Some wearables with limited functions (fitness trackers and smartwatches) do not have an operating system, whereas others need an operating system capable of supporting an ecosystem of apps.

2.4.1 Open-Source Operating Systems

Android Wear: Somewhat confusingly, Android Wear is a wearable OS that is different from the popular Android OS for smartphones [4]. Android Wear was derived from Android to be suitable for smartwatches and is mainly designed to pair smartwatches to work with Android smartphones (although version 2.0 enables Android Wear smartwatches to run native apps without the need for a smartphone nearby). Android Wear is mostly open source but Google adds a proprietary layer of services such as Google Now (for voice recognition).

Android: Android itself is not designed for wearable devices but can be modified for a wearable. Like Android Wear, Android can run on the ARM Cortex-A processor and potentially any processor supporting the Linux kernel (which Android is based on).

Tizen: Tizen is an open-source OS, also based on Linux, started by a group of companies in 2011 as an alternative to Android [65]. While it has not found success in smartphones, it has been adopted for a significant number of smartwatches. Tizen is most commonly found in Samsung smartwatches with ARM Cortex-M processors. Tizen supports apps in the Tizen app store (native apps are written in C, whereas Android apps are written in Java).

Embedded Linux: Wearables may choose embedded Linux because Linux is open source and supported on a wide variety of processors including ARM Cortex-M, Cortex-A, MIPS, and x86. Linux is a general-purpose OS, which means that apps can be developed easily, but Linux might be an overkill for a wearable designed for limited functions.

mbed OS: The open-source mbed OS is based on a real-time operating system, CMSIS-RTOS RTX [8]. Supported by ARM, mbed OS runs on a range of Cortex-M processors. For security, a supervisory kernel called uVisor helps to isolate security domains used to restrict access to memory and peripherals.

2.4.2 Proprietary Operating Systems

watchOS: Apple's watchOS is a version of its proprietary iOS customized for its Apple Watch. The original Apple Watch used a custom S1 system in package (SiP) that integrated an application processor, memory, storage, and support processors for wireless communications, sensors, and I/O controllers in a sealed package. The Apple Watch series 2 uses the S2 SiP. It is known that iOS, and hence watchOS, is based on the XNU kernel, a hybrid between BSD and Mach kernels.

Windows 10: Recently, Microsoft designed the latest Windows 10 to work across the broadest range of machines including wearables. One of its central features is the so-called "universal app" platform which means that developers can create apps that will run across all Windows devices of any size and form factor.

Pebble OS: Pebble OS was an operating system developed for the Pebble smartwatch until Pebble Technology was shut down in December 2016. Pebble OS was based on the FreeRTOS kernel, a real-time OS for embedded devices. Most of Pebble's intellectual property and staff, except hardware, were purchased by Fitbit.

LinkIt OS: MediaTek offers a proprietary LinkIt OS specialized for the Aster SoC which features low power and low cost [37]. It has a low-power standby mode, which enables always-on wearable devices to have small energy footprints. The battery life of devices can reportedly last a few days with normal usage.

WebOS: WebOS based on the Linux kernel was originally created by Palm as the successor to Palm OS. Palm was acquired by HP which released the operating system as open source under the name Open webOS. HP licensed webOS to LG Electronics in 2013 for its web-enabled smart TVs. WebOS made it into the LG Watch Urbane LTE but the current line of LG Watch Urbane (with Qualcomm Snapdragon processors) supports only Android.

WearableOS: WearableOS is a special package of the Unison RTOS (real-time operating system) [55]. A real-time OS has a deterministic preemptive kernel and a small memory footprint. WearableOS is specifically designed to minimize power consumption and support a range of sensor and wireless technologies.

3 Wireless Communications Security

Wearables take advantage of a number of wireless technologies to communicate with other devices or a cloud service [25, 32, 45]. As mentioned in Sect. 2.2, wearables have limited processing, memory, and power resources. Wearables can save resources

by sending its sensor data (after preprocessing) to a device with more resources, e.g., a smartphone.

This section gives an overview of the common wireless technologies which differ in several ways: range; data rates; spectrum; error control; robustness against interference, atmospheric attenuation, and various sources of noise; and protection against eavesdropping. The IEEE 802.15 working group has developed a few standards for wireless communications applicable to wearable devices, namely, IEEE 802.15.1 (Bluetooth), IEEE 802.15.4 (Zigbee), and IEEE 802.15.6 (body area networks). Other protocols including ANT+, UWB (Ultra-wideband), NFC (near field communication), IEEE 802.11 (Wi-Fi), GPRS (General Packet Radio Service), and UMTS (Universal Mobile Telecommunications System) are also used among wearable devices. There is not much to say about IEEE 802.11, GPRS, or UMTS because they are general-purpose wireless services not designed particularly for wearables.

Wireless communications are expected to be an avenue for attackers. Wireless communications face the same security risks as wired communications (e.g., eavesdropping, data modification, packet injection, masquerade, and replay) except that attacks are easier to accomplish in the radio environment. For instance, eavesdropping on radio signals is easy for any receiver within range, whereas a wired link requires a physical tap. An unsecured wireless link may expose personal data, or worse, allow an adversary to bypass other security mechanisms, and compromise a wearable device.

As security risks are well known, each wireless technology includes security mechanisms. Cryptography is the foundation for secure communications. The standard encryption algorithm advanced encryption standard (AES) is typically employed to ensure confidentiality. However, protocols may differ in the choice of key length, block cipher mode, method for key agreement (key distribution), and calculation of MAC (message authentication code) for data integrity. Another important difference may be how devices are authenticated to each other.

For wearables, it must be kept in mind that they have very limited computation and power resources. Consequently, traditional cryptographic approaches for encryption and key establishment may not be well suited [67]. For instance, public key cryptography is considered to be too computationally demanding for wearables, and hence private key cryptography is assumed. However, this raises the question of how symmetric keys will be distributed securely.

3.1 Bluetooth

Bluetooth is standardized as IEEE 802.15.1, but the commercial technology is managed by the Bluetooth Special Interest Group (SIG) consisting of more than 30,000 companies [60]. Bluetooth is popular due to its design oriented at simple and low-cost implementation. It is widely implemented in smartphones, fitness trackers, wireless earphones, and other accessories. Bluetooth 4.0 provides a specific stack for

low-power communications called Bluetooth Low Energy (BLE), also marketed as Bluetooth Smart, that is particularly relevant for wearables.

BLE utilizes 40 radio channels with 2 MHz spacing in the 2.4 GHz unlicensed band [27]. BLE communication is divided into two phases: advertising and data communication. Advertising messages use 3 out of the 40 available RF channels and allow device discovery and connection establishment. Once the advertising device (e.g., wearable) receives a connection request from the master device (such as a smartphone), the data transfer phase starts. Both paired devices can start exchanging data frames through the remaining 37 RF channels using adaptive frequency hopping. Communications between paired devices are limited between 10 m and 1 Mbps.

BLE allows one device serving as the master connected with an unlimited number of slaves to form an ad hoc piconet. A slave in one piconet can act as the master for another piconet simultaneously, thus creating a chain of networks called a scatternet.

Due to its wide adoption, Bluetooth security has been studied extensively [12, 21]. Security features include stealth, frequency hopping, authentication, and encryption.

Stealth: Devices can hide and refuse connections through non-discoverable and non-connectable modes. Normally in discoverable mode, devices reply to inquiries, letting other nearby devices discover their existence, but in non-discoverable mode, devices do not announce their presence by ignoring inquiry scans. In connectible mode, devices listen for requests to their Bluetooth address whereas in non-connectible mode, they do not allow other devices to initiate connections.

Frequency hopping: BLE uses frequency hopping spread spectrum (FHSS) to mitigate interference between devices but it helps to protect against eavesdropping. A device follows a pseudorandom sequence to hop among 37 different radio channels that are established during connection establishment [29]. In order to eavesdrop, an adversary has to determine the hopping sequence. Unfortunately, the limitations of BLE connections allow an attacker to easily get the sequence [56].

Authentication: Bluetooth has four security modes for authentication and encryption. The first three (modes 1 to 3) apply to legacy versions, while mode 4 applies to current versions. Security mode 1 is insecure with no authentication or encryption. Mode 2 (service-level enforced security) uses authentication and encryption at the service level, after a channel has been established. Mode 3 (link-level enforced security) uses authentication and encryption at the link-level connection is established. Mode 4 offers secure simple pairing (SSP) to create service-level security, similar to security mode 2.

SSP simplifies the pairing process compared to legacy Bluetooth which uses a personal identification number (PIN) to authenticate devices (not users). In comparison, SSP offers four association models that are flexible in terms of device input/output capability:

- Numeric comparison for a pair of Bluetooth devices capable of displaying a six-digit number and asking the user to enter a yes/no response on each device if the numbers match.
- Passkey entry for one Bluetooth device with input capability (e.g., keyboard) and another device with a display but no input capability.

- Just works where at least one device does not have a display or a keyboard for entering digits (e.g., headset).
- Out-of-band (OOB) for a pair of devices that support a common additional wireless or wired communication channel for device discovery and cryptographic value exchange.

SSP also improves security through the addition of elliptic-curve Diffie–Hellman (ECDH) key agreement to generate a secret symmetric key called long-term key (LTK). ECDH is a variation of the well-known Diffie–Hellman protocol [20] that makes use of elliptic-curve cryptography [35]. The Diffie–Hellman protocol allows two devices to establish a shared secret (in this case, the LTK) by exchanging public numbers over an insecure communication channel. ECDH is believed to be strong against passive eavesdropping and man-in-the-middle (MITM) attacks during pairing.

Each device generates its own ECDH public–private key pair using P-256 or P-192 elliptic curves. Each device sends the public key to the other device according to the Diffie–Hellman protocol. The devices then perform stage 1 authentication which is dependent on the association model (described above).

Bluetooth 4.2 added the secure connections feature which upgraded low-energy pairing to utilize advanced encryption standard—cipher-based message authentication code (AES-CMAC) and P-256 elliptic curve. This means that the LTK is generated based on an AES-CMAC-128 function. Also, when both BLE devices support secure connections, P-256 elliptic curves are used; otherwise, P-192 curves are used during ECDH.

Bluetooth 4.2 renamed low-energy pairing to low-energy legacy pairing. As legacy pairing does not use ECDH, it provides no eavesdropping protection and is considered broken for all pairing methods except OOB.

Encryption: BLE uses advanced encryption standard—counter with cipher block chaining message authentication code (AES-CCM) encryption [68]. AES-128 is a U.S. standard block cipher with 128-bit keys. CCM combines cipher block chaining mode with MAC authentication. The CCM mode generates an encrypted keystream that is applied to input data using the XOR operation and creates a 4-byte MAC in one operation. It is difficult for an eavesdropper to decrypt packets without intercepting packets in the initial key exchange phase.

During pairing, the LTK is generated and stored locally in each device. There is no exchange of the LTK, and therefore, pairing is not vulnerable to interception of the LTK by an eavesdropper. The link is encrypted by AES-CCM using an encryption key derived from the LTK. AES-CCM is used to provide confidentiality as well as per-packet authentication and integrity.

There is no authentication challenge/response step to verify that both devices have the same LTK or CSRK. The LTK is used to generate the link encryption key, and therefore, successful encryption implicitly provides authentication.

Bluetooth 4.0 introduced two features: low-energy private device addresses and data signing. These two features involve the generation of two keys: the identity resolving key (IRK) and connection signature resolving key (CSRK).

If BLE's privacy feature is enabled, the IRK maps a resolvable private address (RPA) to an identity address. The identity address is a static random address or a public address. The IRK allows a trusted device to determine the identity address of another device from an RPA which can be dynamic. Previously, a device would have to be assigned a static public address, and the public address could be learned during discovery. If that device remained discoverable, its location could be tracked by an adversary.

The CSRK is used to verify cryptographically signed attribute protocol (ATT) data frames from a Bluetooth device over unencrypted links. This allows a Bluetooth connection to use data signing (providing integrity and authentication) instead of data encryption (AES-CCM provides confidentiality, integrity, and authentication).

A number of vulnerabilities and attacks specific to Bluetooth are known [21]. These include the following:

- Bluebugging exploits a security flaw in the firmware of some older Bluetooth devices to gain access to the device and its commands.
- Bluesnarfing exploits a firmware flaw in older Bluetooth devices to gain access to the device.
- Bluejacking is an attack similar to phishing that consists of an unsolicited message to convince the user to respond in a certain way or add a new contact to the address book.
- Bluetooth fuzzing consists of malformed data sent to a device's Bluetooth radio and observing how the device reacts.
- Legacy pairing is susceptible to eavesdropping.
- A number of techniques can force a remote device to use Just Works SSP and then exploit its lack of man-in-the-middle protection.

3.2 Zigbee

Based on the IEEE 802.15.4 standard, Zigbee is designed for low-power wireless personal area networks (WPANs). It is intended to offer a simpler and less expensive alternative to Bluetooth or Wi-Fi for applications that do not require a high data rate (i.e., up to 250 kbps). It operates in 16 channels, each 2 MHz bandwidth, that are 5 MHz apart in the 2.4 GHz unlicensed band. It can also use regional unlicensed bands: 784 MHz in China, 868 MHz in Europe, and 915 MHz in the USA and Australia.

Commercialization is overseen by the Zigbee Alliance [2], which publishes application profiles to support interoperability between different products. Also, the alliance certifies Zigbee devices that meet power, bandwidth, and battery requirements. For instance, Zigbee devices should have a minimum battery life of 2 years and output radio power of 0–20 dBm (1–100 mW). For its low power and low data rates, the main applications of Zigbee include wireless sensor networks, embedded sensing, medical data collection, smoke and intruder warning, and building automation. However, it has not been popular for wearables so far.

Zigbee is flexible in terms of supporting star, tree, and mesh network topologies. In each topology, one node acts as a coordinator, including creation of the network. The central node in a star network must be the coordinator. The tree and mesh topologies are useful for transmitting data long distances by multi-hopping through devices acting as Zigbee routers.

The Zigbee RF4CE specification defines a low-cost communications standard that is able to provide reliable levels of connectivity for consumer electronics. It was specifically designed for applications requiring simple device-to-device control communications that do not need the full-featured mesh networking capabilities offered by Zigbee. RF4CE reduces memory size requirements and the cost of implementation. Examples of applications anticipated by the Zigbee Alliance include lighting, fan control, garage door openers, and keyless entry systems. Its purported advantages include channel agility using three channels instead of 16, a power management mechanism for all device classes, a discovery mechanism for nodes, multiple star topology, inter-PAN communication, and a security key generation mechanism.

Building on the basic security framework defined in IEEE 802.15.4, Zigbee implements most security procedures at the network and application layers, which cover key establishment, key transport, frame protection, and device management. Security is based on the AES-128 encryption cipher. Several suites combining AES-128 and MACs of various lengths are offered with increasing security levels as follows:

- no security;
- confidentiality only: AES-CTR (AES-128 in counter mode);
- authentication only: AES-CBC-MAC (AES-128 cipher block chaining message authentication code) with 32-, 64-, or 128-bit MAC;
- confidentiality and authentication: AES-CCM (same as BLE described above) with 32-, 64-, or 128-bit MAC.

A 128-bit key can be associated with either a network or a link. An initial master key must be obtained through a secure medium (transport or preinstallation). The security of the entire network depends on the master key. Link keys are derived from the master key. Link and master keys are only visible to the application layer. Various services use different one-way variations of the link key to avoid security risks.

Zigbee authentication is performed using ECMQV (elliptical curve Menzies–Qu–Vanstone), a key agreement protocol based on Diffie–Hellman using elliptic curves. It is believed to be a secure form of authentication.

One special device that is trusted by the other devices is recognized as the trust center. The trust center keeps the network key and provides point-to-point security. Ideally, devices will have the trust center address and initial master key preloaded. The trust center provides a network key to typical applications that do not have special security needs.

Many attacks on Zigbee have been investigated. Physical attacks include malicious signal interference; Zigbee can change frequency channels in the presence of interference, but it is relatively slow (Zigbee does not use frequency hopping). Physical access to a Zigbee device's RAM may access the encryption key which is often

flashed on all the devices in a Zigbee network. An adversary may be able to use a special serial interface on a Zigbee device to capture the encryption keys as those keys are moved from flash to RAM during power up.

Encryption keys might be captured remotely. Zigbee radios use pre-shared keys or over-the-air (OTA) key delivery. OTA delivery may be attacked by a malicious node mimicking a node on the Zigbee network to capture packets, which can then be analyzed and decrypted using free and open-source equipment.

Replay and/or injection attacks may be able to trick Zigbee devices into performing unauthorized actions. Zigbee devices are susceptible to these types of attacks because of the lightweight design of the protocol, which has very minimal replay protection and session checking.

3.3 IEEE 802.15.6

The IEEE 802.15.6 standard specifies communications for a type of WPAN called wireless body area network (WBAN) to interconnect low-power devices that are implanted within the body or mounted on the body. WBAN is limited to a short range within the immediate proximity of a human body. A WBAN might utilize a WPAN device as a gateway to the Internet.

In order to support a variety of medical, consumer, and entertainment applications, the standard includes three physical layers: narrowband, UWB (ultra-wideband), and HBC (human body communication) in frequency bands around 400 MHz, 800 MHz, 900 MHz, and 2.4 GHz.

Three levels of security are prescribed in IEEE 802.15.6 [67]:

- level 0 unsecured communications: data frames have no encryption, data authentication, or integrity assurance;
- level 1 authentication only: frames use authentication but not encryption;
- level 2 authentication and encryption: data frames use authentication and encryption.

One of the security levels is selected during the association process where a node and a hub identify themselves to each other. A master key (MK) is established between them for unicast secured communication or a pre-shared key is activated. A pairwise temporal key (PTK) is created for each new session. For multicast secured communication, a group temporal key (GTK) is shared with the corresponding group using the unicast method.

A 256-bit key establishment is based on the Diffie–Hellman protocol with elliptic curves. The cipher-based message authentication code (CMAC) is used to derive the MK and key message authentication codes (KMAC). Initially, the node and hub have a pre-shared MK. The node initiates the association process by sending a security association frame request. The hub responds by joining, and the pre-shared MK is activated and shared between the node and hub by mutual agreement. Then, a new PTK is generated and shared.

Data frames can be transmitted in secured or unsecured communication modes. Nodes that do not require security receive all frames including beacons without validating the security information. The secured frames are authenticated and encrypted or decrypted using the AES-128 CCM mode (as in Zigbee and BLE).

3.4 ANT+

ANT is a proprietary ultralow-power protocol for wireless sensor networks from ANT Wireless, owned by Garmin [5]. It is similar to BLE but oriented toward applications with sensors. Communication range is limited to 20 m, and data rate is low (bursts up to 60 kbit/s) in the 2.4 GHz band. ANT can be used for body area networks, personal area networks, and local area networks.

ANT+ is an interoperability function added to the base ANT protocol to allow nearby ANT+ devices to work together to collect sensor data. ANT+ uses "device profiles" that specify how data is transmitted between devices, including the data format, channel parameters, and other communication parameters. For example, ANT+ enabled fitness monitoring devices such as heart rate monitors, pedometers, speed monitors, and weight scales can all work together to assemble and track performance metrics. Device profiles are shared among all ANT+ adopters, enabling any ANT+ adopter to create an interoperable device.

As a proprietary WSN protocol, not much is known about ANT+ security except that it is based on keys. ANT+ network keys are required to access the ANT+ network. Network keys are generated and provided by the ANT Alliance. Only devices with the same profiles and network keys can communicate with each other. Network keys must be requested from the ANT+ Alliance, an open special interest group of companies, after subscribing to be an ANT+ adopter.

3.5 UWB

Similar to spread spectrum, UWB spreads data across a very wide spectrum, in this case, defined to be at least 500 MHz of spectrum or 20% or more of the center frequency. As a result, the power spectral density is very low which limits the interference with conventional radio systems using the same spectrum. In the U.S., the federal communications commission (FCC) approved UWB in the 3.1–10.6 GHz range at a power level of −41.3 dBm/MHz or 75 nW. The spectrum above 3 GHz avoids overlap with GPS, cellular, and many other services.

UWB was appealing for short-range, high data rate applications but suffered a couple of setbacks. First, the IEEE 802.15.3a task group attempted to bridge competing UWB proposals from the UWB Forum and the WiMedia Alliance. The IEEE 802.15.3a task group was deadlocked for several years and eventually disbanded in 2006. Most vendors went with the WiMedia Alliance specifications using

orthogonal frequency division multiplexing (OFDM). The specification divides the allowed spectrum into 528 MHz sub-bands of OFDM channels. Data rates can reach 480 Mbps at a range up to 10 m.

The second problem was competition from other high-speed wireless technologies being standardized by the IEEE 802.11 working group. In 2009, IEEE 802.11n offered a maximum single-channel data rate exceeding 100 Mbps and a theoretical maximum overall data rate of 600 Mbps using 40-MHz bandwidth with four spatial streams. Then IEEE 802.11ac, an extension of 802.11n, offered a single-link minimum of 500 Mbps and overall 1 Gbps in the 5 GHz band. IEEE and the wireless gigabit alliance (WiGig) jointly developed IEEE 802.11ad offering short-range theoretical speeds up to 7 Gbps in the 60 GHz unlicensed band. However, 802.11ad requires substantial power and is limited to line of sight. UWB also has advantages in greater resistance to noise, superior security, high jamming resistance, greater multipath immunity, low-power consumption, and high-penetration ability.

As a physical layer technology, most security issues handled in higher protocol layers are not relevant to UWB. The main security threat is eavesdropping. Because of the low average transmission power, UWB has an inherent immunity to detection and eavesdropping. An eavesdropper has to be very close to the transmitter (about 1 m) to be able to detect transmissions. In addition, UWB pulses are time modulated with codes unique to each transmitter/receiver pair. The time modulation of extremely narrow pulses adds more security to UWB transmission, because detecting picosecond pulses without knowing when they will arrive is nearly impossible.

Naturally, data will be encrypted but there is a question of whether standard encryption algorithms such as AES may consume too much power. It has been proposed to save power by pushing part of the cryptography into the physical layer by hiding the signal in the time domain [34]. The transmitter and receiver share a secret key. The key is used to randomly offset UWB pulses such that an eavesdropper cannot detect the signal coherently without knowing the key.

3.6 NFC

NFC is for short-range wireless communications (limited to 10 cm) commonly used for contactless payments. It is also used for sharing photos and files between devices, and enabling devices to act as identity authentication, e.g., keycards. Two NFC devices within 10 cm use electromagnetic induction between antennas to exchange data up to 424 kbps in the 13.56 MHz unlicensed band. As a fairly low-rate but easy-to-use technology, NFC is also useful to set up more capable wireless connections such as Bluetooth.

NFC is covered by a number of standards starting from earlier ones on radio frequency identification (RFID): ISO/IEC 14443, FeliCa (by Sony), ECMA-340, ECMA-352, ISO/IEC 21481, and ISO/IEC 18092. The NFC Forum promotes implementation and standardization of NFC technology to ensure interoperability between devices and services [22].

In comparison with BLE, NFC has advantages of much lower cost and easier set up (versus pairing between BLE devices), but NFC suffers from a much shorter range and lower data rate.

NFC is an option for BLE out-of-band key exchange in addition to being a viable communication technology itself. Obviously, the short communication range is one natural challenge for eavesdroppers [26, 30]. The radio signal for wireless data transfer might be picked up less than 10 m, depending on multiple parameters. Also, passive devices are much harder to eavesdrop on than active devices, and an eavesdropper may have to be within a few centimeters.

However, plain NFC does not ensure secure communications and various attacks have been demonstrated. There is no protection against eavesdropping, data modification, or man-in-the-middle attacks. Applications use higher layer cryptographic protocols (e.g., SSL/TLS) for security.

4 Device Security

Wearables are vulnerable to attacks on hardware and software like any other computing devices.

4.1 System Security

Conventional desktop computers and operating systems such as Windows and Mac OS X are loaded with security features such as trusted platform module (TPM) chip, hard drive encryption, secure protocol suites (e.g., SSL/TLS and SSH), code signing, sandboxing, anti-malware software, and built-in firewalls. In comparison, wearable devices have much less computing, memory, and power resources, which impose serious limitations on feasible security features.

As mentioned earlier, wearables use embedded processors and SoCs. More security features are being implemented in these processors such as the TrustZone technology in the ARM Cortex-M23, Cortex-M33, and Cortex-A, and the secure enclave in Apple's A-series SoCs [7].

Wearable operating systems are a broad mixture of open-source (mostly based on Linux) and proprietary operating systems, with varying capabilities. Linux is a widely used operating system that is generally believed to be fairly secure. It is difficult to ascertain the security of proprietary operating systems.

Traditional cryptography poses a challenge for wearables. There is a recognized need for new lightweight cryptographic solutions with countermeasures to side-channel attacks that will be better for resource-constrained wearables [17].

4.2 Vulnerabilities

Verifying the firmware at update time is a step toward securing IoT devices; however, this is often done by the onboard software that is trusted to be authentic [7]. The implementation of this check must be sound. For example, schemes that utilize random numbers must ensure the usage of a cryptographically secure random number generator, and any used cryptographic certificates must be validated by a trusted certificate authority.

It may not be sufficient to just authenticate updates [7]. The software stack should also be authenticated; otherwise, the validity of an update cannot be determined reliably. Also, a proper chain of trust in the hardware architecture is needed before authenticating the software stack.

If a device is remotely updated, it must be able to check the integrity and authenticity of downloaded updates [7]. Typically, updates are protected cryptographically. However, errors and vulnerabilities have been seen in implementations.

Another point of vulnerability is debug interfaces [7]. Circuit board must expose programming interfaces and test points for testing the different components on the board. These interfaces are not removed after testing and might be used by adversaries to inject malicious code.

In sophisticated wearables capable of running different apps, there is a risk that apps might have vulnerabilities exploitable by adversaries. Since wearable apps are designed with tight hardware constraints, these apps can be inherently weaker than apps developed for desktop computing. For example, runtime bound checking might be eliminated to save computational power and memory space, thus exposing the apps to buffer overflow attacks.

In desktop computers, exploits might be caught and blocked by a host-based intrusion detection system (IDS). However, wearables do not have the computation and power resources to run intrusion detection [19].

A number of studies have experimentally looked for vulnerabilities in various commercial fitness trackers [28, 50, 70]. Most of the vulnerabilities found were related to the insecure implementation of communication protocols.

4.3 Malware

Much like desktop computers, wearables will be targets for malware [7]. Wearables are attractive targets because they hold a considerable amount of valuable information. Moreover, they are always connected to the network.

Linux has seen malware such as the Mirai bot. Some security companies anticipate an increase in Linux malware caused by an expansion of Internet of Things devices.

If wearables have any protection, it might consist of software level solutions such as firmware signing and code signing. Wearables do not have sufficient resources for traditional anti-malware software.

Hardware Trojans may also pose a threat. These are malicious modifications to integrated circuits that are difficult to detect by normal testing methodologies because they might be subtle. For example, a hardware Trojan inserted into a SoC might weaken the entropy of the random number generator used to generate keys. If these keys are used for encryption, the computational effort required by an adversary to decrypt data could be reduced greatly [7]. Hardware Trojans could require expensive specialized tests to detect them.

5 Privacy Issues

Most people think of privacy as the problem of data exposed to an eavesdropper, which is solved by encryption. However, data may be exposed in various ways. Privacy is a broader problem of a user controlling every aspect of where his or her personal data is represented, sent, stored, accessed, and possibly deleted.

Wearable devices including fitness trackers and medical devices are capable of collecting a variety of sensitive personal data. Therefore, they may be subject to privacy and regulatory policies such as the Health Insurance Portability and Accountability Act (HIPAA) that states the obligation for companies operating in the US to protect healthcare information [1].

Privacy issues are real for commercial wearable devices. An investigation of several wearable fitness trackers found a number of general privacy concerns [46].

5.1 Access Controls

Access control works in enforcing different access rights for different users. Data should be classified based on the sensitivity and each user will have different access levels. For example, a doctor will have more access rights than a nurse. Access control consists of authenticating the user, granting appropriate privileges, and revoking privileges. Due to their hardware constraints, the implementation of access control in wearables is still an open issue.

Examples of hardware implementations of protected data include ARM TrustZone, Apple's secure enclaves, and Samsung KNOX, as discussed earlier.

Access to data stored in the cloud must also be designed carefully. Homomorphic encryption has been proposed to ensure confidentiality of sensitive health data in the cloud [48]. Caregivers might be able to analyze the data which is unreadable to others, including the cloud service provider. However, homomorphic encryption is not practical for resource-constrained wearables.

5.2 Outsourcing

The current generation of wearables is much better than similar past devices (e.g., pedometers) in terms of their seamless integration with cloud services and online social networks. This integration raises security and privacy issues because by design, social networks are inherently open.

Unfortunately, wearables are too resource constrained to perform conventional methods to protect health data, e.g., de-identify data by data aggregation or removing common identifiers. An alternative is to move data to the cloud to take advantages of the computing and storage resources of cloud services. However, this approach introduces other privacy issues that have not been worked out entirely [70].

5.3 Health-Related Information of Non-health-Related Applications

Although most fitness trackers and wearable devices are not marketed as medical devices and therefore are not covered by health data protection regulations, they do store a considerable amount of user data that could be derived to extract health-related information. A study of BLE data traffic between a fitness tracker and a smartphone found a correlation to the intensity of the user's activity [18]. Experimental results with the Fitbit app and tracker showed that when the app was opened, the tracker would send a different amount of BLE packets depending on the activity the user was performing. This means that simply by observing and analyzing the encrypted BLE packets, an adversary could be able to guess the user's current activity (walking, sitting, and running).

5.4 Tracking

Wearable devices become particularly useful when they are connected to other devices. When wearable devices are interconnected, there could be a continuous exchange of data among them without being noticed by humans. In such a scenario, privacy may be easily breached (e.g., by revealing locations) [19].

This risk to privacy has been observed, for instance, in the Jawbone tracker. The BLE specification recommends that devices should change their Bluetooth device address frequently in order to prevent tracking, but this privacy feature is not implemented in the Jawbone tracker. It always uses the same address [28]. The static address allows the user to be tracked across visited locations.

In a similar way, Fitbit and other fitness trackers are constantly advertising themselves irrespective of whether they are already paired with some device or not

[18, 28]. In this case, the tracker uses the same device address and does not change it despite the BLE guidelines. Thus, a user might be tracked by listening to the Bluetooth traffic in an area.

6 Conclusions

Wearables are a diverse and expanding class of computing devices that pose many security and privacy issues, but our experience with them as a mass consumer device is limited to the past few years. The issues are more challenging for two major reasons. Wearables are designed to collect, store, and share a great deal of health data that might be considered personal or sensitive. At the same time, wearables have limited computation, memory, and power resources to implement a full suite of security features.

The increasing popularity of wearables among consumers is pushing commercial wearables into the marketplace before security and privacy issues can be worked out. This chapter has highlighted these issues because more research is needed to incorporate security features into wearables from the beginning of their design.

References

1. Al Alkeem, E., Yeun, C.Y., Zemerly, M.J.: Security and privacy framework for ubiquitous healthcare IoT devices. In: 10th International Conference for Internet Technology and Secured Transctions (ICITST), pp. 70–75. IEEE (2015)
2. Alliance, Z.: Zigbee alliance. http://www.zigbee.org
3. Amorim, V.J.P., Delabrida, S., Oliveira, R.A.R.: A constraint-driven assessment of operating systems for wearable devices. In: VI Brazilian Symposium on Computing Systems Engineering (SBESC), pp. 150–155. IEEE (2016)
4. Android: Android wear. https://www.android.com/intl/en_uk/wear/
5. ANT: The wireless sensor network solution this is ant. https://www.thisisant.com
6. Aoyagi, T., Miyasaka, K.: Pulse oximetry: its invention, contribution to medicine, and future tasks. Anesth. Analg. **94**(1 Suppl), S1 (2002)
7. Arias, O., Wurm, J., Hoang, K., Jin, Y.: Privacy and security in Internet of Things and wearable devices. IEEE Trans. Multi-Scale Comput. Syst. **1**(2), 99–109 (2015)
8. ARMmbed: mbed os. https://www.mbed.com/en/platform/mbed-os/
9. Baldini, F.: Invasive sensors in medicine. In: Optical Chemical Sensors, pp. 417–435. Springer (2006)
10. Blasco, J., Chen, T.M., Tapiador, J., Peris-Lopez, P.: A survey of wearable biometric recognition systems. ACM Comput. Surv. (CSUR) **49**(3), 43:1–35 (2016)
11. Boll, S., Pulsipher, D.: Suppression of acoustic noise in speech using two microphone adaptive noise cancellation. IEEE Trans. Acoust. Speech Signal Process. **28**(6), 752–753 (1980)
12. Bouhenguel, R., Mahgoub, I., Ilyas, M.: Bluetooth security in wearable computing applications. In: International Symposium on High Capacity Optical Networks and Enabling Technologies, pp. 182–186. IEEE (2008)
13. Braasch, M.S., Van Dierendonck, A.J.: Gps receiver architectures and measurements. Proc. IEEE **87**(1), 48–64 (1999)

14. Catalano, J.T.: Guide to ECG analysis. Lippincott Williams & Wilkins (2002)
15. Chatterjee, A., Aceves, A., Dungca, R., Flores, H., Giddens, K.: Classification of wearable computing: a survey of electronic assistive technology and future design. In: 2nd International Conference on Research in Computational Intellegence and Communication Networks (ICRCICN), pp. 22–27. IEEE (2017)
16. Cornacchia, M., Ozcan, K., Zheng, Y., Velipasalar, S.: A survey on activity detection and classification using wearable sensors. IEEE Sens. J. **17**(2), 386–403 (2017)
17. Cruz, R.J., Reis, T.B., Aranha, D.F., Kupwade Patil, H.: Lightweight cryptography on arm. In: NIST Lightweight Cryptography Workshop. NIST (2016)
18. Das, A.K., Pathak, P.H., Chuah, C.N., Mohapatra, P.: Uncovering privacy leakage in ble network traffic of wearable fitness trackers. In: 17th International Workshop on Mobile Computing Systems and Application, HotMobile '16, pp. 99–104. ACM (2016)
19. Di Pietro, R., Mancini, L.V.: Security and privacy issues of handheld and wearable wireless devices. Commun. ACM **46**(9), 74–79 (2003)
20. Diffie, W., Hellman, M.: New directions in cryptography. IEEE Trans. Inf. Theory **22**(6), 644–654 (1976)
21. Dunning, J.P.: Taming the blue beast: a survey of bluetooth based threats. IEEE Secur. Priv. **8**(2), 20–27 (2010)
22. Forum, N.: Nfc forum. http://nfc-forum.org
23. Gao, W., Emaminejad, S., Nyein, H.Y.Y., Challa, S., Chen, K., Peck, A., Fahad, H.M., Ota, H., Shiraki, H., Kiriya, D.: Fully integrated wearable sensor arrays for multiplexed in situ perspiration analysis. Nature **529**(7587), 509–514 (2016)
24. Gartner: Gartner says worldwide wearable devices sales to grow 18.4 percent in 2016 (2016). http://www.gartner.com/newsroom/id/3198018
25. Ghamari, M., Arora, H., Sherratt, R.S., Harwin, W.: Comparison of low-power wireless communication technologies for wearable health-monitoring applications. In: 2015 International Conference on Computer, Communication, and Control Technology (I4CT), pp. 1–6. IEEE (2015)
26. Ghosh, S., Goswami, J., Kumar, A., Majumder, A.: Issues in NFC as a form of contactless communication: a comprehensive survey. In: International Conference on Smart Technology and Management for Computing, Communication, Controls, Energy and Materials (ICSTM), pp. 245–252. IEEE (2015)
27. Gomez, C., Oller, J., Paradells, J.: Overview and evaluation of bluetooth low energy: an emerging low-power wireless technology. Sensors **12**(9), 11734–11753 (2012)
28. Goyal, R., Dragoni, N., Spognardi, A.: Mind the tracker you wear: a security analysis of wearable health trackers. In: 31st Annual ACM Symposium on Applied Computing (SAC '16), pp. 131–136. ACM (2016)
29. Gupta, N.K.: Inside Bluetooth Low Energy. Artech House (2016)
30. Haselsteiner, E., Breitfuß, K.: Security in near field communication (NFC). In: Workshop on RFID security, pp. 12–14 (2006)
31. Islam, N., Want, R.: Smartphones: past, present, and future. IEEE Pervasive Comput. **13**(4), 89–92 (2014)
32. Islam, S.K., Fathy, A., Wang, Y., Kuhn, M., Mahfouz, M.: Hassle-free vitals. IEEE Microw. Mag. **15**(7), S25–S33 (2014)
33. Jiang, H., Chen, X., Zhang, S., Zhang, X., Kong, W., Zhang, T.: Software for wearable devices: challenges and opportunities. In: IEEE 39th Annual Computer Software and Applications Conference, pp. 592–597. IEEE (2015)
34. Ko, M., Goeckel, D.L.: Wireless physical-layer security performance of UWB systems. In: IEEE MILCOM 2010, pp. 2143–2148. IEEE (2010)
35. Koblitz, N.: Elliptic curve cryptosystems. Math. Comput. **48**, 203–209 (1987)
36. Labrador, M.A., Yejas, O.D.L.: Human Activity Recognition Using Wearable Sensors and Smartphones. Taylor and Francis Group, Boca Raton, FL (2014)
37. Labs, M.: Linkit assist 2502. https://labs.mediatek.com/en/platform/linkit-assist-2502

38. Lara, O.D., Labrador, M.A.: A survey on human activity recognition using wearable sensors. IEEE Commun. Surv. Tutor. **15**(3), 1192–1209 (2013)
39. Ltd, A.: Arm processors. http://www.arm.com/products/processors
40. Ltd, A.: A system-wide approach to security. https://www.arm.com/products/security-on-arm/trustzone
41. Moody, G., Mark, R., Bump, M., Weinstein, J., Berman, A., Mietus, J., Goldberger, A.: Clinical validation of ecg-derived respiration (edr) technique. Comput. Cardiol. **13**, 507–510 (1986)
42. Mosenia, A., Sur-Kolay, S., Raghunathan, A., Jha, N.K.: Wearable medical sensor-based system design: a survey. IEEE Trans. Multi-Scale Comput. Syst. **PP**(99), 1–1 (2017)
43. Moyer, J., Wilson, D., Finkelstein, I., Wong, B., Potts, R.: Correlation between sweat glucose and blood glucose in subjects with diabetes. Diabetes Technol. Ther. **14**(5), 398–402 (2012)
44. Nourbakhsh, N., Wang, Y., Chen, F., Calvo, R.A.: Using galvanic skin response for cognitive load measurement in arithmetic and reading tasks. In: 24th Australian Computer-Human Interaction Conference, pp. 420–423. ACM (2012)
45. Pantelopoulos, A., Bourbaki, N.G.: A survey on wearable sensor-based systems for health monitoring and prognosis. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **40**(1), 1–12 (2010)
46. Paul, G., Irvine, J.: Privacy implications of wearable health devices. In: 7th International Conference on Security of Infomation and Networks (SIN '14), pp. 117:117–121. ACM (2014)
47. Perera, C., Liu, C.H., Jayawardena, S.: The emerging Internet of Things marketplace from an industrial perspective: a survey. IEEE Trans. Emerg. Topics Comput. **3**(4), 585–598 (2015)
48. Preuveneers, D., Joosen, W.: Privacy-enabled remote health monitoring applications for resource constrained wearable devices. In: 31st Annual ACM Symposium on Applied Computing (SAC '16), pp. 119–124. ACM (2016)
49. Qualcomm: Snapdragon wear 2100 processor product brief. https://www.qualcomm.com/documents/snapdragon-wear-2100-processor-product-brief
50. Rahman, M., Carbunar, B., Topkara, U.: Secure management of low power fitness trackers. IEEE Trans. Mob. Comput. **15**(2), 447–459 (2016)
51. Ray, S., Park, J., Bhunia, S.: Wearables, implants, and Internet of Things: the technology needs in the evolving landscap. IEEE Trans. Multi-Scale Comput. Syst. **2**(2), 123–128 (2016)
52. Roggen, D., Perez, D.G., Fukumoto, M., van Laerhoven, K.: Iswc 2013—wearables are here to stay. IEEE Pervasive Comput. **13**(1), 14–18 (2014)
53. Roman, R., Najera, P., Lopez, J.: Securing the Internet of Things. IEEE Comput. **44**(9), 51–58 (2011)
54. Romano, S.M., Pistolesi, M.: Assessment of cardiac output from systemic arterial pressure in humans. Crit. Care Med. **30**(8), 1834–1841 (2002)
55. RoweBots: Wearableos. https://rowebots.com/en/products/wearableos-unison-wearable-operating-system
56. Ryan, M., et al.: Bluetooth: with low energy comes low security. In: WOOT (2013)
57. Salvo, P., Francesco, F.D., Constanzo, D., Ferrari, C., Trivella, M.G., Rossi, D.D.: A wearable sensor for measuring sweat rate. IEEE Sens. J. **10**(10), 1557–1558 (2010)
58. Samsung: Exynos 9 series (8895). http://www.samsung.com/semiconductor/minisite/Exynos/w/solution/mod_ap/8895/
59. Sazonov, E., Neuman, M.R.: Wearable Sensors: Fundamentals. Academic Press, Implementation and Applications (2014)
60. SIG, B.: Bluetooth technology website. https://www.bluetooth.com
61. Starner, T.: The challenges of wearable computing: Part 1. IEEE Micro **21**(4), 44–52 (2001)
62. Starner, T.: How wearables worked their way into the mainstream. IEEE Pervasive Comput. **13**(4), 10–15 (2014)
63. Statista: Wearable device shipments worldwide from 2015 to 2021 (in million units). https://www.statista.com/statistics/610478/wearable-device-shipments-worldwide/
64. Tamura, T., Maeda, Y., Sekine, M., Yoshida, M.: Wearable photoplethysmographic sensors—past and present. Electronics **3**(2), 282–302 (2014)
65. Tizen: Tizen. https://www.tizen.org

66. Vashist, S.K.: Non-invasive glucose monitoring technology in diabetes management: a review. Anal. Chim. Acta **750**, 16–27 (2012)
67. Wang, S., Bie, R., Zhao, F., Zhang, N., Cheng, X., Choi, H.A.: Security in wearable communications. IEEE Netw. **30**(5), 61–67 (2016)
68. Whiting, D., Housley, R., Ferguson, N.: Counter with cbc-mac (ccm). In: RFC 3610. IETF (2003). http://www.ietf.org/rfc/rfc3610.txt
69. Zhang, M., Raghunathan, A., Jha, N.K.: Trustworthiness of medical devices and body area networks. Proc. IEEE **102**(8), 1174–1188 (2014)
70. Zhou, W., Piramuthu, S.: Security/privacy of wearable fitness tracking iot devices. In: 9th Iberian Conference on Infomation Systems and Technologies (CISTI) (2014)
71. Zieniewicz, M.J., Johnson, D.C., Wong, D.C., Flatt, J.D.: The evolution of army wearable computers. IEEE Pervasive Comput. **1**(4), 30–40 (2002)

Wearable Computing and Human-Centricity

Arash Tadayon, Ramin Tadayon, Troy McDaniel and Sethuraman Panchanathan

Abstract Wearable computing has gained significant traction with the advancement of computing technology and the resulting increase in efficiency and power within smaller form factors. The interfaces and applications of these devices have evolved over centuries from very primitive implementations to those which adapt to and anticipate user needs. However, the principles of human-centric design have only recently been defined and understood in the context of wearable computers. These devices introduce an additional set of requirements to the traditional concepts of human-centric computers which have yet to be defined in an adaptable framework. The failures of many recently launched wearable devices highlight the importance of these considerations throughout the design and development process. Human-centricity currently serves as one of the major challenges to the ubiquity and future success of wearable devices.

1 Introduction

Although the topic of wearables has recently gained immense popularity with the advancement of mobile technologies, the history of wearable technology spans well over four centuries. The earliest examples of wearables date back to the 1500s with the invention of the first timepieces designed to be worn on an individual. The invention and distribution of wearable computers, however, began several 100 years later within the twentieth century. Since then, these devices have grown in parallel with computational power and have become increasingly ubiquitous in their applications.

A. Tadayon · R. Tadayon · T. McDaniel (✉) · S. Panchanathan
Arizona State University, 699 S. Mill Ave., Tempe, AZ 85281, USA
e-mail: troy.mcdaniel@asu.edu

A. Tadayon
e-mail: arash.tadayon@asu.edu

R. Tadayon
e-mail: ramin.tadayon@asu.edu

S. Panchanathan
e-mail: panch@asu.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_12

Their applications range from fashionable accessories to, more notably, medical devices.

Similarly, the methods of interaction for these devices, including sensing and feedback, have diversified. Traditional wearables relied on visual or haptic interfaces to gather information from users; however, new technologies allow for more discreet methods of input including, for example, electrical stimulation. These passive methods of data gathering have introduced a new paradigm of anticipatory interfaces in wearables that anticipate user needs. This diversification and evolution of wearables has lead to their development in industries ranging from fashion to assistive devices.

While wearable devices are increasingly integrated as commonplace technology within society, their ultimate role remains uncertain. Many organizations continue to invest in the future of wearable technology, but with varying degrees of success; most notably, the Google Glass project was intended to transform human perceptions of wearables and their potential, but was ultimately deemed unacceptable by the general public.

The evolution of wearable devices has also introduced a new consideration in the design of technology: human-centricity. Society is mandating design that takes into account the needs of the individual and addresses these needs throughout the ideation and development of new devices. Wearables have introduced a mobile context to computing that offers new restrictions on the function and appearance of devices in order to best support the individual without inhibiting external needs. To adapt to these changes, many developers have introduced personal, social, cultural, and environmental considerations to the decision-making process in the design of new technology. New interaction paradigms have emerged which consider the users internal and external context in the delivery of information.

2 Definitions

The following definitions serve to clarify and disambiguate several of the terms used in this chapter. The main concepts of "Human-centricity" and "Wearable Computing" are defined in their respective sections below.

Mobile Computing: We define Mobile Computing, or "Nomadic Computing" [71], as the design, implementation, and usage of portable devices which can access and transmit information without requiring a fixed location. Mobile devices such as smartphones and laptops are the central focus of the study of mobile computing. This definition includes mobile wearables which travel with the user.

Universal Design: The term "universal design" was originally coined by [43] in relation to the construction of buildings. It was intended to describe the process of designing all products and the built environment to be esthetic and usable to the greatest extent possible by everyone, regardless of their age, ability, or status in life.

Assistive Technology: We adhere to the definition of Assistive Technology outlined in the Assistive Technology Act of 2004 (29 U.S.C. Sec 2202(2)): "any item,

piece of equipment, or product system, whether acquired commercially, modified, or customized, that is used to increase, maintain, or improve functional capabilities of individuals with disabilities" [2].

Usability: We rely on the definition of "usability" set forth by the International Organization for Standardization as the "extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."

Human Factors: While there are various definitions for this term in recent literature, in this chapter, we follow James H. Stramler's definition of Human Factors as the "field which is involved in conducting research regarding human psychological, social, physical, and biological characteristics, maintaining the information obtained from that research, and working to apply that information with respect to the design, operation, or use of products or systems for optimizing human performance, health, safety, and/or habitability" [73].

Accessibility: We refer to accessibility as the design of products to include access for people with disabilities. This design strategy can facilitate either direct access to the individual or indirect access through assistive technologies.

Anticipatory Interface: Robert Rosen defines an anticipatory system as one that "contains an internal, predictive model of itself and its environment, which allows it to change state at an instant in accord with the models predictions pertaining to a later instant" [70]. Such an interface takes a passive approach to prediction, but ultimately operates in the domain of action in response to predictions of the users context.

Modality/Multimodality: We refer to a "modality" in the sense commonly used in the study of Human–Computer Interaction (HCI): it is a perceptual channel through which information is transmitted between a human and a machine, or between two human beings [33]. Multimodality consequently refers to the use of multiple modalities, as defined above, in a system or interface.

Smart Device: A "smart device" can be defined as a multifunctional ubiquitous device which is able to communicate, often wirelessly, with other devices, access remote and locally stored information, and provide access to that information to the user via a mobile user interface [65]. Examples of these devices include smartphones, smartwatches, and modern tablets.

3 What Is a Wearable Computer?

Since it is the primary focus of this chapter, we begin our discussion of wearable computing with a definition of the concept:

Wearable computer is a broad term used to describe any computer that is worn to some degree on or inside a human's body. Due to the wide scope of devices that this term can encompass, it is more beneficial to characterize it rather than use an explicit definition. In 1997, Rhodes described a wearable computer as having five main characteristics [69]:

- **Portable while operational**: The most distinguishing feature of a wearable is that it can be used while walking or otherwise moving around. This distinguishes wearables from both desktop and laptop computers.
- **Hands-free use**: Military and industrial applications for wearables especially emphasize their hands-free aspect, and concentrate on speech input and heads-up display or voice output. Other wearables might also use chording keyboards, dials, and joysticks to minimize the use of a user's hands.
- **Sensors**: In addition to user inputs, a wearable should have sensors for the physical environment. Such sensors might include wireless communications, GPS, cameras, or microphones.
- **Proactive**: A wearable should be able to convey information to its user even when not actively being used. For example, when a new email arrives, your computer should be able to notify you immediately of its arrival.
- **Always on, always running**: By default, a wearable is always on and working, sensing, and acting. This can be contrasted to pen-based PDAs, which normally sit idle in one's pocket and are only activated when being actively used for a task.

Since these characteristics address form factor, input/sensing, feedback/delivery of information, and operational aspects of devices, they operate within the modern definition of a computer.

4 History of Wearable Computers

The concept of wearable computing dates back to the 1500s, with the invention of wearable timepieces that were transitional in size between clocks and watches. These clock-watches were designed to be worn as jewelry on clothing or around the neck and utilized only an hour hand. Although their calculations were fairly imprecise, and they are not considered computers in the modern sense, these were the first wearable devices that computed time.

These primitive wearables were followed by the development of rings that served as a fully functional abacus in the 1600s and the invention of the wristwatch in the 1800s. However, the first generation of modern wearable computers had not emerged until the twentieth century. For this reason, we begin our timeline in this section at the twentieth century with the development of wearable computing devices.

The invention of the first modern wearable computer was self-credited to Edward Thorp and Claude Shannon for their device which aided in predicting the outcome of a roulette wheel in 1960 [76]. The device worked by measuring the position and velocity of the ball and rotor to predict their future paths and stopping points. It was concealed within a user's shoe and used radio transmission to inform another individual of the winning number as shown in Fig. 1.

The 1970s and 1980s yielded the emergence of general purpose wearable computers, and the release of the first wearable computers built for general consumers. Hewlett-Packard released the HP-01, the first algebraic calculator watch, at this

Fig. 1 Thorp and Shannon's shoe

Fig. 2 HP-01 watch

time [45]. The watch was released with six interactive functions: time, alarm, timer/stopwatch, date/calendar, calculator, and memory. It included 28 keys, 6 of which were operated by finger and the remainder through a stylus fitted in the watch-band clasp (Fig. 2).

The 1990s and 2000s ushered in the integration of new sensors, such as cameras, into wearables enabling new applications including augmented reality. These decades also introduced the first applications of wearables which were implanted in the human body. In 2002, as a part of his Project Cyborg, Kevin Warwick implanted an electrode array to measure electrical signals within his nervous system and relay the information to a pendant worn by his wife that would change colors based on

the data [77]. The late 2000s also saw the introduction of mobile phones which were integrated into wristwatches. Due to the diversification of wearables and the introduction of Personal Area Networks (PANs) and Body Area Networks (BANs), the need for standardization of interfaces and communication between these devices had become apparent. Thus, the Institute of Electrical and Electronics Engineers (IEEE) and Internet Engineering Task Force (IETC) began to develop standards for communication protocols such as Bluetooth.

The influence of wearables on technologys evolution was solidified in the 2010s as many companies began developing their own wearable devices, further expanding their ubiquity. A suite of exercise bands (Nike Fuel Band, Fitbit, JawBone, etc.), smart watches (Pebble, Apple Watch, and Galaxy Gear), and assistive devices (exoskeletons, prosthetics, etc.) were rapidly released and adopted by the public within this relatively short timeframe. There were, however, some failures within this relentless expansion. Most notably, Google released the Google Glass as a head-mounted optical display in an effort to introduce seamless augmented reality to the average individual. The release of this device introduced a myriad of privacy and safety concerns ranging from the ability to discreetly record to the usage of the device while operating a motor vehicle. The technology was ultimately discontinued within a year of its launch.

As indicated above, hundreds of years of history have served to shape the purpose of wearables and their integration with modern ubiquitous computing technology. This history has indicated that wearable technology is subject to the evolution and advancement of society and individuals. The recent technological advances above set the stage for an ongoing discussion of the ability of these devices to facilitate variation in user needs and backgrounds, and the challenges and barriers introduced by these factors.

5 Wearable Device Interfaces

In this section, we review the design of wearable interfaces and the advantages and limitations that various design strategies place on user interaction. Furthermore, we introduce some of the primary challenges in interface design for modern wearables, particularly when considering the individual user.

One of the main barriers of entry for wearable devices is the burden of adaptation. When a new device is deployed, users are often forced to adapt to a new interface and method of interaction to adequately use the technology. This has traditionally been one of the main challenges for new devices, which developers have attempted to address by defining a universal set of human factors design considerations. New interaction styles are often derived from existing methods of interaction to minimize adaptation costs.

Similarly, due to the inherent contextual challenges associated with wearable devices, interfaces have an additional set of considerations beyond traditional Human–Computer Interaction that they must fulfill to be considered user-friendly [12].

There are three main considerations outside of the scope of traditional user experience:

Interaction Period: These devices are often used within much smaller interaction periods than conventional devices, and are thus required to be highly efficient with respect to user attention.
Context: They are often used in dual-task contexts where the user is simultaneously performing some other task while interacting with the device. The primary task is often some physical task in the real world and thus, interaction with the wearable becomes almost a distraction [78].
Interface Simplicity: When considering the broad range and volume of devices developed in this domain, it is important that the interfaces are not overly complex and map to interactions that build on what a user is already familiar with. Although interface simplicity is a consideration made in broader applications, the importance of this consideration is much greater in the wearable domain because of a combination of the above factors.

Thus, the main goal of wearable user interfaces is to support users during their day-to-day tasks while minimizing cognitive load and interaction time. These devices typically incorporate multimodal feedback in user interaction. They often rely on more than one sense to relay information back to the user; however, one of their feedback channels typically serves as the primary communication modality with secondary channels that are often redundant. Effectiveness of feedback channels is not only highly dependent on the individual, but also on current context, and it can change as the user's environment varies. As an example, a visual interface on a smartwatch might be a good way to alert a user that they are receiving a phone call while sitting idle, but may not be as effective when driving. In this scenario, haptic displays may prove more useful to users while driving since they do not shift the users vision from his or her primary motor task. Incorporating multimodality allows users to prioritize which sense they want to dedicate to the reception of information without completely disrupting their primary task.

5.1 Wearable Interfaces by Modality

The modality of interaction for a wearable interface varies greatly depending on its intended purpose and application. We identify three main modalities of feedback for wearable interfaces: visual, auditory, and haptic. Examples and descriptions of the usage of each modality are provided below.

5.1.1 Visual

Visual interfaces are the most common method that wearables use to relay information to users since more than 70% of our sensory receptors are visual, and engage

Fig. 3 Wearables with visual interfaces. Pebble Smartwatch, FitBit, Google Glass

almost 50% of our cortex [21]. Typically, these devices rely on displays mounted on a user's wrist or head, but may occasionally depend on a non-wearable devices display (for example, smartphones) to communicate wirelessly transmitted information. There is, however, an explicit limitation on the size of these devices and displays due to an individuals visual acuity and their ability to discern minor changes (Fig. 3).

Visual displays can be dated back to the earliest wearable devices as clocks rely on a visual display to relay information on time. With the advancement of technology, these primitive interfaces turned to digital screens with varying degrees of resolution. Most digital displays for wearables can be categorized as either head-mounted or wrist-mounted.

Head-Mounted Displays (HMDs) are wearable, lightweight displays mounted on a user's head and have digital displays in front of at least one of the user's eyes. These displays can be further separated into monocular (only displays to a single eye) and binocular (displays that cover both eyes) with the former being the more recent approach to HMD development [42]. Within these subcategories exist immersive displays that inhibit an individual's ability to perceive the real world outside of the HMD and semi-transparent, non-immersive displays. Applications of immersive and non-immersive HMDs range from aviation, where they are used for navigation and to enhance situational awareness for pilots [74], to computer-aided drafting for model understanding and manipulation [7].

Because they hinder an individual's ability to see their surroundings, immersive displays are not often seen on individuals in day-to-day environments [3]. These devices have, however, found many applications within the field of virtual reality. Devices like the Oculus Rift and the HTC Vive have introduced virtual reality to the gaming community, motivating developers to create applications that enhance the experience of the average consumer. These devices were designed to operate with limited mobility as they obscure vision and fully divert attention from the outer environment, but can still be used in limited mobile environments as they provide virtual representations of the real-world environment. Thus, more recently, HMDs have been geared toward non-immersive displays for use in augmented reality applications outside of virtual environments.

Although non-immersive HMDs have design limitations [63, 79] which include reduced vision due to the veiling luminance of the display, more of today's wearables (outside of virtual reality) are shifting toward this approach as it abides relatively well by one of the basic premises of wearable displays: that the users primary task should not be interrupted [15]. Google Glass was a monocular, non-immersive HMD that aimed to provide contextual information to users through a small projection within a glass in a user's peripheral vision. A broad range of applications within industry and research were explored including a display for pediatric surgeons [52] or high-level activity recognition using blinking and head motion [30].

Wrist-Mounted Displays (WMDs) are digital displays that are worn on the user's wrist or forearm. The position of these displays presents inherent difficulties for a user as it distracts attention from the surrounding environment, and thus, they often inhibit a user's primary motor task. These devices typically require the user to lift his or her arm while interacting with the device, which can lead to issues with muscle fatigue after prolonged use; consequently, although they often provide similar contextual information (for example, navigational or environmental data), WMDs typically have shorter interaction cycles than HMDs and have different interruption techniques. WMDs often rely upon secondary modalities such as haptics to direct the user's attention outside of their visual field.

The main category of WMDs is smartwatches. A smartwatch is defined as "a wrist-worn device with computational power, that can connect to other devices via short-range wireless connectivity; provides alert notifications; collects personal data through a range of sensors and stores them; and has an integrated clock" [8]. Although they have existed in many different forms for decades, smartwatches took off with the launch of the Pebble device. This was the first platform-agnostic watch that allowed users to interact with their smartphones without having to take the device out of their pockets. Although smartwatches have started to gain interest, they still do not offer enough additional functionality when compared with smartphones to allow for mass adoption [8].

5.1.2 Auditory

The second most popular form of feedback occurs through the auditory channel. An auditory display is defined as "any method of communicating information, usually non-textual information, by means of nonspeech sound" [72]. This channel offers the opportunity to receive information while the eyes and hands may be busy performing some other primary task. Auditory interfaces can be subdivided into three main categories: **verbal**, **audification** and **sonification** [72].

Verbal: Verbal interfaces use natural speech to present information. Examples include car navigation systems, hands-free smartphone assistants, and automated museum tours. These interfaces are the most common in wearable devices.
Audification: This technique is a direct mapping for data points into audio signals to produce sound patterns. It is limited in use to large, periodic data sets where

patterns can be explored at a high level rather than looking for granular differences. An example of this would be looking through large sets of financial data to compare trends between years and playing tones that might be higher in pitch to denote higher profits and lower pitches to denote lower profits or losses.

Sonification: This is an analogic approach to mapping data to sound. Frequency, harmonicity, and pulse rate are used as variables in the development of associations between sound and data. This allows for more fine-grained exploration of data since mappings can be made to specific characteristics. As an example, a Geiger counter uses the rate of clicking to denote the radiation level in the environment.

Sound is often used in wearable interfaces to offload visual information. This is vital because, as discussed in the previous section, visual displays on wearables are often small and constrained. Care must be taken in the amount of information presented visually to avoid problems of cognitive overload and confusing, cluttered displays. Thus, audio cues are often used to represent a subset of the information that may be presented through a nonvisual channel. Sound can reliably attract a user's attention while they perform another task and can, therefore, serve as a method to interrupt the user and redirect his or her attention to a visual display.

5.1.3 Haptic

Haptic interfaces rely on an individual's sense of touch to provide information and are used both as a primary and secondary modality for feedback in wearable interfaces. Haptic interfaces "generate a feedback to the skin and muscles, including a sense of touch, weight, and rigidity" [31]. Due to restrictions on size, haptic stimulation is most often accomplished through vibrotactile patterns in wearable devices. As more wearable devices are developed, the exploration of haptics becomes more of a necessity than a luxury since the visual and auditory channels are primarily used during navigation. This leaves haptics as an unobstructed channel that can be used without severely impacting day-to-day activities but still allowing the user to receive information from their device.

Because the sense of touch lacks the spatial acuity of vision and the temporal acuity of hearing, frameworks based on natural human speech have been proposed as building blocks for information delivery using haptics [48]. As noted previously, the most common example is that of vibrating smartwatches to inform a user that he or she is receiving a notification or call. Applications of haptic interfaces for wearables range from navigation where a 4-by-4 array of micromotors have been used on a user's back to present directional information [16] to emotional therapy where human touch can be recorded and played back [5]. While tactile displays are most commonly used in conjunction with visual or auditory displays, new devices are starting to explore the value of haptic-only wearables (e.g., Moment) (Fig. 4).

The true benefit of haptic interfaces lies not only in the small space required for actuators but also in their capability to be personal and discreet. These interfaces have an advantage that is unlike any other interface in that a user can receive information

Fig. 4 Moment smartwatch with haptic interface (https://wearmoment.com/)

and interact with their device without alerting those around them to the interaction. This allows haptics-enabled wearable devices to seamlessly integrate into social contexts and augment rather than interrupt.

5.2 Examples of Modern Wearable Interfaces

In the expanse of wearable applications, interfaces for these devices have developed unique characteristics of interaction. Two primary examples illustrating the potential of these characteristics, anticipatory and invisible interfaces, are described below. Although these examples do not comprise the entirety of modern wearable interfaces, they are mentioned here as they have gained significant attention in recent research.

5.2.1 Anticipatory Interfaces

With the rise of mobile devices, we are able to infer more about a user's location, activity, and social setting than ever before. As these devices continue to advance with new sensors, we may see a shift from inference to prediction of context that will, in parallel, open the door for anticipation within computing applications. Although the principle of anticipation has been known, most existing approaches on the interaction cycle for assistive devices have been "laissez-faire." Put simply, the device will wait for explicit interaction from the user before it processes and provides an output.

In discussing applications, it is important to first differentiate prediction from anticipation since the two are often incorrectly used interchangeably. Predictive applications are those that simply build predictions of the user's current or future

context. Anticipation uses these predictions to impact the future to the benefit of the user. Applications in the field of anticipatory computing rely on two key steps prior to the ability to anticipate. These include sensing the surrounding context and creating a predictive model of this context. Once the predictive model has been created, the system then uses this for anticipation of a user's future needs [61].

Because the concept of anticipation in wearable computing is so new, few applications exist that are truly anticipatory. The majority of work in this area involves robotics. Within robotics, the principle of anticipation has taken the forefront in navigation [22], perception [27], and human movement characterization [23, 68]. Similarly, authors have explored applications in gaming through eye tracking to predict a player's actions [35]. These early systems have helped to show the applicability and usefulness of anticipation, but are restricted in context.

Within the wearable domain, there is an abundance of recent literature surrounding predictive applications of internal and external context. One example of an application that emphasizes the usefulness of mobile phones in determining external context is SoundSense [41]. This project explores the use of the microphone to determine information such as activity, location, and social events. The authors proposed a scalable framework for modeling sound events and were able to classify four different activities: walking, driving a car, riding an elevator, or riding a bus (the precision on riding a bus was much lower than other conditions). Similarly, a project that explored internal context, called EmotionSense, was designed to infer a user's emotional state from microphone data [67].

Furthermore, one of the other major efforts toward the classification of human behavior and extrapolation of context through mobile phones is Darwin Phones [50]. The authors developed the first framework in the mobile domain which could automate the updating of models over time, pool models that have been created and evolved within other mobile devices, and combine classification results from multiple mobile phones. This methodology is a step above most other work in the literature that relies on the local sensing abilities of a single mobile device rather than crowd-sourcing the classification.

Pejovic and Musolesi have proposed the potential for applications of anticipatory computing within the emerging field of digital behavior change interventions in mobile environments [61, 62]. The authors have referenced UbiFit [13] and BeWell [36] as two applications that have taken very rudimentary steps toward the inclusion of anticipation in mobile applications and provided potential architectures for applications in this domain [61, 62]. UbiFit is a personal health application designed to monitor weekly activity and provide subtle feedback when users are not active enough. The app displays a garden which thrives when the user is meeting activity goals, and remains barren when inactivity persists. BeWell is a mobile application that monitors a user's health along three dimensions: sleep, physical activity, and social interaction. Much like UbiFit, this application provides intelligent feedback to the user to promote better health through an ambient display of an aquatic background which becomes more active the healthier you are.

5.2.2 Invisible Interfaces

A crucial factor impacting the future expansion of wearable devices is ease of use. With the examples stated above, the authors have embedded the interface for the applications into the existing interface of the mobile phone to create an unobtrusive feedback loop. Pantic et al. take this process one step further and state that the key to anticipatory interfaces is "ease of use" and the ability to "unobtrusively sense certain behavioral cues of the users and to adapt automatically to his or hers typical behavioral patterns and the context in which he or she acts" [58].

It is this ability to unobtrusively sense behavior cues and to use those as inputs for technology that comprises an invisible interface. Essentially, the traditional methods of explicit human–computer interaction is abstracted away from the user and instead use both internal and external context as the primary inputs for the technology. This promotes the principles of ubiquitous computing and makes the technology an extension of the person.

Although most applications in mobile computing still maintain the need for interaction with a physical interface, there has been a major effort toward the development of context-aware applications [10, 64]. These systems are generally split into two major subcategories: external context (physical) and internal context (logical). Context is any information that can be used to characterize the situation of an entity where an entity is a person, place, or object that is considered relevant to the interaction between a user and an application, including location, time, activities, and the preferences of each entity [28]. These systems face many challenges including determining relevant information, dealing with uncertainty, and privacy [4].

The literature in this domain is quite developed and uses context as an input in a variety of different ways. One unique approach looks to combine the influences of internal and external user context to proactively determine recommendations [39]. The system builds a context history and a profile for the user in each of these contexts to accurately predict the user's needs simply based on past behavior.

Fenza et al. explore the usefulness of internal context in the healthcare domain by using a network of wearable sensors to determine the individual's current state of health and provide personalized services. The authors use Fuzzy Logic to automatically characterize context and find healthcare services that approximately meet this context [17].

Muoz et al. explored the development of a context-aware messaging system in a hospital environment [53]. Users (doctors, nurses, physicians, etc.) were given mobile devices to write messages to each other that are only sent when a specified context is encountered. For instance, a nurse could leave a message for the next doctor entering a given room. The system automates the delivery based upon sensed context across many devices. However, this system does not fully embrace the concept of an invisible interface since the main method of interaction is still physical rather than an automated interaction solely based on context.

6 Areas of Application

As the baby boomer population continues to age, the adoption of wearable, monitoring technology grows at a considerable rate. New wearable devices are coming out everyday that explore applications across all facets of life ranging from entertainment to health. Health has specifically become a focal point in this progression due to the exponentially-growing demand for these devices. The last few years have introduced a broad variety of exercise bands (Nike Fuel Band, Fitbit, JawBone, etc.), smart watches (Pebble, Apple Watch, and Galaxy Gear), and mobile health apps. These technologies offer access to information that was previously unavailable and provide platforms for a myriad of new assistive technologies through remote monitoring.

6.1 Fashion

One historically significant concern in the adoption rate and usability of wearable technology is whether society considers technology to be "fashionable." This is a cultural concern dating as far back as the introduction of eyeglasses and watches, and is an integral part of the introduction of wearables to human societies. One of the most important elements of fashion is that it relies heavily on context: different cultures, societies, and regions have different takes on what makes a piece of clothing or accessory "fashionable." Even within the United States, for example, fashion interests can vary greatly by subregion [26]. With this as the basis for our understanding of fashion, we present a list of integration strategies for the design of fashion-aware wearable technology:

Assimilation: Some wearable technologies, particularly "electronic textiles", can be embedded or woven into, or made to resemble, existing fashionable clothing or accessories in their region of deployment. This strategy favors users who value discretion in the technology they use by hiding the circuitry and interface of the wearable device, essentially rendering the technology "invisible" on the wearer. This strategy can be seen, for example, in Liu et al.'s e-textile pants for stability assessment in the elderly with motion impairments [38].

Enhancement/Augmentation: In this strategy, wearables are made to "enhance" the appearance of clothing or accessories either by attaching themselves atop these items or by being embedded in such a way that their existence is obvious, either through exposed circuitry or through an exposed interface (Fig. 5). Often these devices are adopted by users who value the high-tech look and wish for their wearables to stand out. Mistry and Maes' SixthSense gestural device, for example, uses a worn pendant to project an interface to augment the real world [51].

Separation: The final strategy involves designing devices so that they can be worn separately from clothing and other accessories, while remaining fashionable. This is perhaps the most difficult strategy, as it involves designing a wearable that does not conform to the form factor of common accessories and articles of clothing

Fig. 5 Conductive thread used to create embedded electronics within clothing

in a particular society, but can be considered "fashionable" within that society or culture. Often this means introducing a new category of fashionable items into a society, which can take time to integrate. One example of this is work on fingertip haptic wearables [66].

Regardless of the integration strategy used, there are several general points to consider when designing fashionable wearables. One is that customizability and adaptability in the look and feel of a device can help improve its fashion awareness, as it can then be molded, either manually or autonomously, to match fashion tastes for a variety of users and cultures. Another is that the device should be usable in a variety of different contexts. For example, glasses can be worn to aid users with visual impairment (eyeglasses), improve visual clarity in bright environments (sunglasses), or simply to enhance one's look (fashion glasses or clear glasses). The more functions a wearable can serve, the greater its targeted audience, and the greater the likelihood that it can be adopted into the fashion of a particular society.

6.2 Behavior Modification

One popular use for wearable devices, particularly in healthcare, is the modification of problematic behavior. Devices intended for behavior modification can sense and respond to targeted patterns of behavior to promote positive outcomes for the user. We can classify these devices into two main subcategories: **facilitators** and **drivers** [60]. **Facilitators** of behavior change afford greater control to a user over changing their behavior. They may remind the user that a problematic behavior is occurring, and provide steps and options for correction. These devices are intended to inform, but not to directly elicit behavior change. On the other hand, **drivers** of behavior change are devices which take direct action to change a user's behavior, often by modifying/constraining the environment. A driver for eye health may, for example,

turn off and disable usage of a television screen once it has detected that the user has been watching for a prolonged period in unhealthy conditions.
There are several general requirements which are commonplace for wearable devices aimed at behavior change:

1. **The device should be able to accurately detect problematic behavior**. This can often be a nontrivial issue, as the portable nature of wearable devices places limitations on what information they can sense in real-time. As a result, some types of behavior are easier to detect and interpret than others. For example, physical activity is a highly studied sensing category for behavior modification devices, and many wearables exist today which can discern this behavior from quantitative indicators such as step count or heart rate [6]. The accelerometer in the average smartphone, for example, can be utilized to estimate step count and get a generally reliable measure for an individual's physical activity, under the assumption that the phone travels with its intended user [80]. However, sleep patterns may be harder to detect, as the mechanisms for automatically detecting sleep patterns without user entry can be complex [11].
2. **The device should be aware of the context in which behavior occurs**. Often the environment and the user's goals play a large role as predictors of certain behaviors. A device intended for behavior modification should be aware of these factors to prevent the occurrence of false positives and false negatives in the detection of problematic behavior [34]. As an example, consider a device intended to detect and correct problematic gait patterns in users with Parkinson's disease. Such a device would detect when Freezing-of-Gait (FoG) events occur as the user walks, as these are dangerous symptoms of the disease given that FoG episodes increase the risk of falling [75] (Fig. 6). However, in a crowded environment, a user may freeze his or her gait simply because the path ahead is blocked or the individual is waiting in a line. These FoG events are not attributable to Parkinson's and should not be treated as problematic behavior. The device would, therefore, need to discern between the different causes for frozen gait and should respond to each appropriately.
3. **The device should provide corrective feedback in a way that is intuitive, accessible, and clear**. It should be immediately apparent to the user, based on the feedback of the device, what problematic behavior is occurring and, in some cases, how to fix it. Furthermore, the feedback given from the device should not produce any unwanted interference that could affect the safety, comfort or health of the user. For wearables, this often means that the device should not be a distraction when walking, driving, or interacting in public situations [18].

Behavior modification devices often base their evaluation on the very same metrics they use to detect problematic behavior. If the problematic behavior can be accurately detected and quantified, then it follows that researchers can evaluate the effectiveness of a device based on the *change* in this problematic behavior produced by usage of the device [54]. The most successful wearables for behavior modification are often able to produce significant change in the targeted behavior under a variety of conditions including users, environments, and contexts.

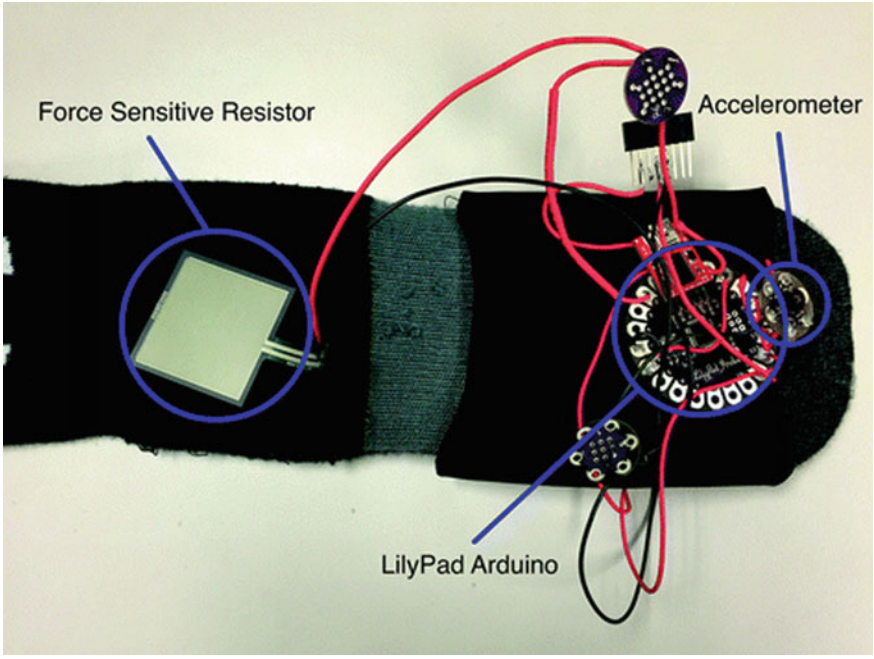


Fig. 6 Device used to track Parkinsonian gait [75]

6.3 Fitness

Where devices for behavior modification are aimed at eliminating or changing problematic behavior, some devices instead focus on helping users maintain positive, healthy lifestyles. By far the most popular of such wearable devices are fitness trackers. These wearables are often facilitators as defined in the above section; they provide the information a user needs to maintain healthy levels of activity throughout the week. As these are devices intended for tracking physical activity, they require a greater attention to durability, comfort, and reliability in design [49]. Inevitably, a wearable for physical activity will be subject to a high amount of movement and shaking, outdoor weather, and potential impact against various surfaces. To account for this, wearable trackers such as accelerometers and heart rates sensors are often well reinforced and insulated against damage to critical circuitry while maintaining as little extra weight as possible to avoid burdening the user.

Since the mid-2000s, the popularity of wearable fitness tracking devices has taken off (Fig. 7). A vast range of wearable devices have been developed from chest straps to shoe attachments that have started to gather more information than ever on an individuals activity and health. The effectiveness of these devices is vastly attributed to the benefits that the smartphone has provided as a central communication system

Fig. 7 Wearable fitness trackers. Jawbone UP, FitBit, Microsoft Band

for this otherwise fragmented market. The industry is now looking at embedding fitness trackers into clothing as well to get even more information on overall health.

6.4 Assistive Devices

Assistive devices are designed, as defined above, to help augment the functional capabilities of individuals with disabilities. A wearable assistive device serves the added benefit of following the user, providing services in many aspects of that user's daily life. One example of a wearable assistive device is the Haptic Belt [47], which can express nonverbal cues to assist in communication for individuals who are blind (Fig. 8). This device uses a pinhole camera that is embedded into a pair of sunglasses to determine if a person is approaching the visually impaired user. The belt then vibrates to allow the user to turn and face the person so that they can initiate conversation. Assistive devices often target a specific disability and a specific goal or function, and can assist in overcoming the challenges related to that function caused by the disability through various means including sensory substitution or augmentation. However, devices that are designed for individuals with disabilities can often have benefits for the population at large as well. As an example, a project called the Note-Taker was developed as a solution for visually impaired students to be able to more accurately take notes in the classroom environment [25]. The device had a camera that communicated with a tablet to allow the student to record the lecture and zoom in on the board to more easily see what was being written. The project had a lot of success with students who were visually impaired but also saw a huge demand from their sighted counterparts who could also benefit from its features.

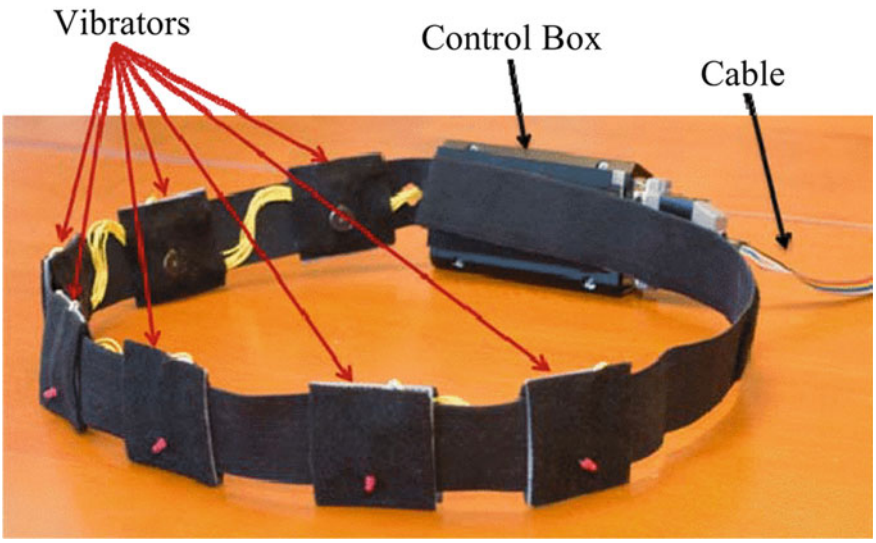


Fig. 8 Haptic Belt device [47]

6.5 Navigation

The problem of navigation often varies wildly by context; as such, wearables intended to assist with navigation can have drastically different requirements and considerations in their design (Fig. 9). To understand the constraints for effective assistance in navigation, we consider a few of the attributes involved in research in this field:

Attributes of the Person:

- **Impairment**: One of the initial considerations in design focus on the user's impairment. Does the user have any attributes that make it challenging to navigate in an environment? Typically, these include sensory, cognitive and/or motor impairments, among others. Often these considerations manifest themselves in the design of the interface, although core functionality may be affected as well. For example, users with visual impairment may need more detailed information about their immediate surroundings [29].
- **Degree of Usage**: Some users rely on navigation assistance more than others. An interface designed for heavy usage may require additional steps to increase battery life, particularly if the device is providing real-time assistance during navigation.

Attributes of the Task:

- **Type of Navigation Task**: A primary concern related to the navigation task is the type of navigation. Is the user walking or driving? What type of vehicle is being used? Each type of navigation task includes its own concerns for safety and responsiveness. For tasks such as driving, the system may need to respond to

Fig. 9 Navigation on the Apple Watch

changes more quickly than for walking due to the relative difference in speed of motion. Walking interfaces may provide the user with real-time information about landmarks along the route [46] while navigation for driving might focus on points of interest at the user's destination [40].

- **Degree of On-the-fly Assistance**: Wearables for navigation may be designed differently depending on how much the device knows about a navigation route beforehand. Devices with high storage capabilities but low real-time memory might pre-calculate optimum routes and follow a predetermined plan for the user's navigation while those with little storage may rely instead on using real-time learning and adaptation to develop a route in real-time for the user, with only short-term planning involved. These types of devices are often more flexible to change in the environment but may use more power or suffer from signal interference. Many navigation algorithms for autonomous agents in AI research, such as that of Oriolo et al. [57], often deal with unknown environments and real-time planning.
- **Path Attributes**: Devices for navigation may also be concerned with details of the path being calculated. For example, in some contexts, the shortest route is desirable while in others a more scenic route is preferable. Furthermore, there may be milestones or checkpoints along the way to a destination that the system should account for in the production of a path.

Attributes of the Environment:

- **Obstructions and Lighting**: In general, navigation devices can improve in the quality of their assistance based on the amount of details of the environment makes available. This includes lighting attributes, which help the system determine which paths are most visible to the system and the user so that it can recommend

the safest route to a destination, particularly at night. For navigation by blind and visually impaired users, information on obstructions and moving objects in the environment may also assist the system in preventing a collision in real-time. The system by Mann et al., for example, utilizes real-time detection by a Kinect camera for collision avoidance [44].

- **Scope/Scale of Navigation Environment**: Finally, navigation systems may also be concerned with the scale of the navigation environment, as it may impact the type of services offered by the system and the number of available routes. Indoor navigation systems such as the one proposed by Golding and Lesh [20] may require higher accuracy of location and orientation detection of the user than navigation within a city.

7 Key Barriers to the Success of Wearable Computers

Although the wearable market is growing at an alarming rate, barriers exist to the full adoption of these devices. We have already seen this with the failure of the Google Glass in 2015 and the fall of the Pebble smartwatch in 2016, but what are the factors that inhibit a wearable technology's success? We identify four main categories of barriers to the success of these devices:

1. **Cost**: Cost has been one of the most important considerations in the rise and fall of wearable technologies. With the rise in popularity of wearable technologies, the markets for these devices mature a lot quicker than traditional technologies. As a result, two things occur: the bottom line prices are driven down and the demand for innovation goes up. This cuts the margins on newer technologies and often leaves only a couple frontrunners. In the case of smartwatches, the Pebble, FitBit, and the Apple watch took over. In order to keep up, these devices must innovate much faster not only on the technical side but also on the manufacturing side. It is the innovations in manufacturing that lead to a decrease in retail price and make the products more affordable for lower income brackets.
2. **Specialization**: Too many highly specialized wearables are being designed with single use cases. This is a barrier since there is an explicit limit (body real estate) on the amount of devices that a single individual can wear at the same time which, if devices are highly specialized, forces users to make decisions of priority on what needs are most important. A survey conducted of user wearing habits indicated that the preferred body parts for various types of wearables are as follows (in descending order): (i) eyes (approximately 72%)—sunglasses, shades, and prescription glasses; (ii) head (approximately 70%)—hats, caps, and scarves; and (iii) hand—wrist watches (68.1%), bracelets (49.7%), and rings (59.4%). Audio earphones and headphones, wearables with which consumers are already familiar, received a preference rate of 64.7% [9].
3. **Social Acceptability**: Form factor and methods of interaction have been major drivers for the social acceptance of wearable computers. Because wearable

devices are still relatively new, more emphasis is currently being placed on the innovativeness of the technology rather than the external appearance and social impacts. As a primary example, Bluetooth headsets were designed to be a modernization for the interaction between people and their smartphones. Iterations of headsets smaller in size began to appear on the market, but ultimately did not achieve mass adoption due to the social awkwardness they created. The devices were so inconspicuous that passersby would not know if the individual was talking to them or talking to someone on their phone. Similarly, the Google Glass' form factor was deemed pretentious and scoffed in public settings. It also introduced questions about privacy that were deemed socially unacceptable. Other factors such as cultural and ethical considerations have an impact on the acceptability of a wearable device within a social setting. For example, in countries which censor social media interactions, devices which augment or enable social media usage by individuals may be deemed unacceptable.

4. **Human-centricity**: There is a set of human factors that are explicit to wearable computers which cannot be ignored. These devices are often used in contexts that are completely unique to those of traditional Human–Computer Interactions. Thus, special considerations need to be made into how these devices are used, when they should interrupt the user, where on the body they should be placed and what their intended purpose is [9].

8 What Is Human-Centricity?

Interfaces are typically categorized as "user-friendly," "accessible," or "intuitive." Human-centricity is a newer design paradigm that aims to address aspects of each of these terms. The term "human-centered" or "user-centered" was coined by Donald Norman in the 1980s [56], and was expanded upon several years later to include four basic suggestions on design [32, 55]:

1. Make it easy to determine what actions are possible at any moment.
2. Make things visible, including the conceptual model of the system, the alternative actions, and the results of actions.
3. Make it easy to evaluate the current state of the system.
4. Follow natural mappings between intentions and the required actions; between actions and the resulting effect; and between the information that is visible and the interpretation of the system state.

The basic principle of human-centric design is that the end user is considered in all stages of design and development. Considerations are made to the needs, wants, limitations, uses, benefits and risks of a device from ideation all the way through development and even marketing [24]. In essence, this philosophy looks to alter the traditional **feedback loop** that exists between designers and developers to also include the class of end users. That representation must be actively involved in the

entire process and constantly giving feedback on their desires since they are an important stakeholder in the final product.

As the concept has matured, standardizations have been developed around human-centricity providing more concrete requirements. Most notably, the International Organization for Standardization created **ISO 13407**: "Human-centred design processes for interactive systems" in 1999. It was revised with **IISO 9241-210**: "Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems" in 2010 and was reconfirmed in 2015 [1]. It describes human-centered design as "an approach to systems design and development that aims to make interactive systems more usable by focusing on the use of the system and applying human factors/ergonomics and usability knowledge and techniques."

9 Concerns of Human-Centricity

ISO 9241-210: "Ergonomics of human-system interaction—Part 210: Human-centred design for interactive systems" recommends six characteristics for human-centered design [14, 19]:

- The adoption of multidisciplinary skills and perspectives.
- Explicit understanding of users, tasks and environments.
- User-centered evaluation driven/refined design.
- Consideration of the whole user experience.
- Involvement of users throughout design and development.
- Iterative process.

These characteristics outline four basic categories of considerations that need to be made in the design of human-centric devices: personal, social, cultural, and environmental (Fig. 10).

9.1 Personal

This is the most critical consideration in human-centric design. The device should be designed in a way that it is aware of the user's limitations from a human factors perspective. The device should be aware of the user's psychological, social, physical, and biological characteristics rather than require the user to attempt to adapt to the device. An example of awareness of physical characteristics would be that if a device is worn on the skin, it should not overheat and potentially burn the user. Similarly, it should support psychological characteristics such as cognitive load. Developing a device with an overly complicated interface that is unintuitive for the intended audience is not human-centric. All of these characteristics vary between individuals and so the device should also consider the spectrum of ability of its intended audience, and address those users through accessibility characteristics.

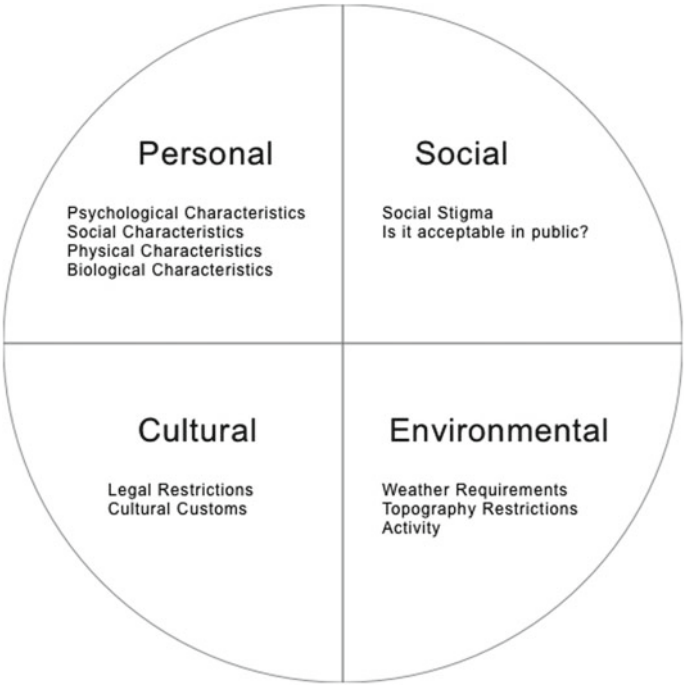


Fig. 10 Consideration categories of human-centric devices

This consideration also dictates that the device should support users in achieving their goals or tasks rather than inhibit them. These tasks might not be directly related to the use of the device but may be done in parallel. As an example, if a mobile, wearable device is designed for active use throughout the day, it should be lightweight so that it does not inhibit mobility as an individual completes his or her day-to-day tasks. The concept of secondary tasks that are not tied with the direct use of the device is often overlooked and can cause a lot of issues with utility and adoption.

9.2 Social

Typically, when thinking about human-centricity, the focus is placed on the considerations made by the individual themselves. However, there are many factors outside of the individual's immediate control that can also dictate design decisions in human-centric devices. Social considerations are one of the most critical as humans are inherently social creatures by nature. Within the design process, if the device is designed for daily use, the question "is the usage of this device acceptable in public?" should be asked. The development team needs to consider the social lives of its users

and how they will be impacted by the technology. As an example, one of the major barriers to the adoption of assistive technology is the social stigma attached to the use of an assistive device [59]. The families of children with disabilities often choose to adopt assistive technology devices due to the perceived increase in attention and visibility.

A secondary consideration to social stigma is whether the device inhibits an individual's ability to effectively communicate with others. This principle of design has less to do with the external appearance and more to do with the paradigm of interaction. Understanding when to interrupt an individual and when to shift into the background is a key factor in designing socially appropriate devices [15]. As an example, smartphones now have settings that allow the user to set certain do-not-disturb hours where notifications are instantly silenced. This mode can be disengaged when the user decides that they are in a social context where it is acceptable to be interrupted.

9.3 Cultural

Culture at any level, including national, regional, or organizational, can greatly influence success in adoption and implementation of devices [37]. Culture plays an important role in the adoption of technology and mandates standards which often underlie or dictate an individual's personal views. These standards appear both in the form of legal restrictions on technology and occasionally as unspoken norms. In the design of person-centric devices, it is important to consider whether the assumptions being made in design implicitly or explicitly violate any of these customs. An example of this consideration having been overlooked was with the launch of the Google Glass. Although regulations did not exist prior to the device, new laws were developed as the product violated cultural expectations of privacy.

Culture also plays an important role in determining economic factors around a device that can often also dictate its potential for future success. When current cultural trends are taken into account, technology can actually begin to mold the future direction of societal views. Apple has been a company that has successfully created a subculture that will readily adopt their new technology with little discretion for whether or not it fits into current cultural values. This comes with the success that they had in designing devices that have become iconic trends within current culture.

9.4 Environmental

The user's external environment is also a major consideration in human-centric design. External context can dictate anything from form factor to actual features and functionality. Obvious questions include: How resistant is the device to the effects of weather, rigorous human activity, crowded environments, dark or bright environments, or low internet signal strength? However, a more obscure consideration would

be the choice of materials to use in the design process. Consider sending a device to be used in a remote area of the world. It would be human-centric to use materials that are readily available in that part of the world in the case that the device needs immediate repairs on site. Similarly, one should consider the environmental impacts that the device may have on the user's environment after they discard the device.

Outside of the natural environment, there is also attention that needs to be paid to the manmade environment. Indoor and outdoor environments have been largely shaped by buildings, sidewalks and roadways which introduce their own set of requirements on human-centric devices. It is important that the device is contextually aware of the user's surroundings and is designed in a way to adapt to the restrictions set forth by this environment. Returning to the example of the assistive device made for monitoring the gait of individuals with Parkinson's disease, it is important to distinguish between the gait of an individual who is walking on an incline and might naturally be taking shorter steps versus an individual who is walking on a flat surface and may be taking short or shuffled steps due to a Freezing-of-Gait episode. Similarly, our gait differs in indoor environments such as waiting in line at a grocery store where we might only be able to shuffle forward every few minutes.

10 Universal Design Versus Human-Centric Design

The term "human-centric design" is often mixed up with "universal design" and the two are many times incorrectly equated. The terms both define considerations to be made about the end-user during the design process, but have slightly different goals.

Human-centered design attempts to model human interaction, personal, social, and cultural values into the design of an interface. The developers are expected to make informed design decisions with respect to the uses, benefits, and risks associated with the end user. This design ideology takes into account not only the characteristics of the individual using the device but also the context in which the device will be used. This presents a new set of requirements on technology that are completely external to the user, but are still important in the design of devices that he or she will use. Similarly, the consequences of design decisions with respect to these factors must be considered at each stage and factored into the decisions that are made in the development of the device. A device is not truly human-centric if it does not address all of these concerns.

Universal design, on the other hand, focuses on enabling as many individuals as possible to access that interface. The main principle is to design with the intent of usability to the greatest extent possible by everyone, regardless of their age, ability, or status in life. The concept of ability is viewed as a spectrum rather than a binary variable and the goal of this design ideology is to allow for the greatest distribution of access across that spectrum. This can be addressed in various ways, but the most common method of implementing principles of universal design is adaptability. The iPad is one of the best examples of universal design as it includes various settings and features that allow for the device to be used by individuals who may have

visual impairments as well as those who may have auditory impairments. Human-centered design focuses on people's interaction with society and with one another while universal design focuses on their interaction with technology.

A piece of translation software, for example, can be human-centered in that it translates subtle parts of human speech such as idioms, special phrases, metaphors, etc., to match the cultural and stylistic format of another language; however, if it cannot be accessed by individuals who are blind, it would not be universally designed. Similarly, a handheld electronic device may have accessibility options that allow for usability to the majority of its intended population; however, if it is too heavy to comfortably carry by hand, it may not be user-centric.

11 Human-Centric Wearables

To design human-centric wearables, one must be aware of the context in which they are used. Often these devices are taken out into the public, moved around, and may stay on the body for long periods of time. All of these requirements oblige us to study how humans interact with the world to ensure that wearables can be embedded into these interactions. Within the design and development process, it is vital to include the end-user's feedback and validates assumptions. The concerns of human-centricity within this context often rely on the context in which the device is used; some general requirements will be covered in this chapter with the most basic division of scope: external to the user versus internal to the user (Fig. 11).

11.1 External Considerations

There are many external factors associated with the use of a wearable computer that are often overlooked in the design process. Although the user may be willing to accept the technology, occasionally it is their external context that places restrictions. It is crucial to understand these facets as they often have a large impact on the ultimate use and adoption of the device.

As an example, cultural appropriateness is something that is often overlooked during the design process of wearables. Would it be culturally acceptable to wear this device in the environment that it was intended to be used in? Many unsupported assumptions are made about what society is willing to accept as appropriate without ever having conversations with the stakeholders. Had this consideration been validated at an earlier point in the design and development process, the wearable could have been made fashionable.

Another example of an external factor would be the wearable computers interference with public interaction. Is it too loud or distracting? Can it lead to awkward social interactions? The example of the Bluetooth earpiece is appropriate. Some devices were so loud that people around an individual could hear the entire

Fig. 11 Internal versus external considerations of human-centric wearables

conversation from both sides which lead to issues with privacy. Other devices were so small that people in the immediate vicinity were not sure if the individual was talking to them or someone on the phone.

11.2 Internal Considerations

Similarly, there are considerations that need to be made about the internal context of a user in the design of human-centric wearables. It is critical to have a complete understanding of the human factors that surround the intended audience of the wearable and use this understanding as a basis for design decisions. Along with a firm understanding of human factors, it is important to acknowledge the person's goals and tasks and ensure that the wearable device empowers the completion of these tasks. As stated previously, wearables are often used in a dual-task setting where interaction with the device is interrupting some other primary physical task in the real world. Thus, the wearable should not become a major disruption to this primary task or inhibit the individual from completing it.

Internal considerations should also span the spectrum of ability of users. This means that the device should be adaptable for individuals of various levels of ability and can often mean redundant interfaces so that it is left to the user to decide how they want to receive information. Redundancy is an important factor in accessibility and, by transfer, is an important aspect of human-centric design. As an example, a watch may provide a visual display when an alarm is triggered but may also provide vibrotactile pattern for haptic feedback and play a tone for auditory feedback.

12 Conclusion

As the era of modern wearable technology continues to advance, the onus of adaptation continues to shift from the user to the wearable device, necessitating the integration of the human-centric design considerations reviewed in this chapter into the development process for these devices. In a global market and a highly connected world, the considerations of the individual, society, and world population form layers of influence in wearable design which often conflict, requiring that developers achieve a delicate equilibrium in their design decisions. The concepts of customizability and multimodality continue to be explored within this context as the single-purposed wearables of yesterday and today begin to assume an increasing number of roles under continuing advancements in the size and power of sensors and mobile interfaces. Future research in this field will seek to yield design patterns, frameworks, models, and prototypes which can address the many limitations and concerns of human-centricity highlighted above without sacrificing the basic standards of usability, cost-effectiveness, and portability that form the groundwork for the success of this technology.

References

1. ISO 9241-11:1998(en): Ergonomic requirements for office work with visual display terminals (VDTs)—Part 11: Guidance on usability, Dec 2016
2. 1994 US Code : : Title 29—Labor : : Chapter 24—Technology Related Assistance for Individuals With Disabilities : : Sec. 2202—Definitions, Jan 2017
3. Baber, C.: Wearable computers: a human factors review. ResearchGate **13**(2), 123–145 (2001)
4. Baldauf, M., Dustdar, S., Rosenberg, F.: A survey on context-aware systems. Int. J. Ad Hoc Ubiquitous Comput. **2**(4), 263–277 (2007)
5. Bonanni, L., Vaucelle, C., Lieberman, J., Zuckerman, O.: TapTap: a haptic wearable for asynchronous distributed touch therapy. In: CHI '06 Extended Abstracts on Human Factors in Computing Systems, pp. 580–585. ACM, New York, NY, USA (2006)
6. Butte, N.F., Ekelund, U., Westerterp, K.R.: Assessing physical activity using wearable monitors: measures of physical activity. Med. Sci. Sports Exerc. **44**(1 Suppl 1), S5–12 (2012)
7. Butterworth, J., Davidson, A., Hench, S., Olano, M.T.: 3dm: a three dimensional modeler using a head-mounted display. In: Proceedings of the 1992 Symposium on Interactive 3D Graphics, pp. 135–138. ACM, New York, NY, USA (1992)

8. Cecchinato, M.E., Cox, A.L., Bird, J.: Smartwatches: the good, the bad and the ugly? In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 2133–2138. ACM, New York, NY, USA (2015)
9. Chen, C.-Y., Tsai, W.-L.: The key success factors of wearable computing devices: an user-centricity perspective. In: *WHICEB*, p. 50 (2014)
10. Chen, G., Kotz, D., and others: A survey of context-aware mobile computing research. Technical report, Technical Report TR2000-381, Dept. of Computer Science, Dartmouth College (2000)
11. Chen, Z., Lin, M., Chen, F., Lane, N.D., Cardone, G., Wang, R., Li, T., Chen, Y., Choudhury, T., Campbell, A.T.: Unobtrusive sleep monitoring using smartphones. In: 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, pp. 145–152. IEEE (2013)
12. Clark, A., Newman, N., Isaakidis, A., Pagonis, J.: What do we want from a wearable user interface. In: Proceedings of Workshop on Software Engineering for Wearable and Pervasive Computing, p. 3. Citeseer (2000)
13. Consolvo, S., McDonald, D.W., Toscos, T., Chen, M.Y., Froehlich, J., Harrison, B., Klasnja, P., LaMarca, A., LeGrand, L., Libby, R., and others: Activity sensing in the wild: a field trial of ubifit garden. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 1797–1806. ACM (2008)
14. I. DIS. 9241–210: Ergonomics of human system interaction-Part 210: Human-centred design for interactive systems. International Standardization Organization (ISO), Switzerland (2009)
15. Drugge, M., Witt, H., Parnes, P., Synnes, K.: Using the "HotWire" to study interruptions in wearable computing primary tasks. In: 2006 10th IEEE International Symposium on Wearable Computers, pp. 37–44, Oct 2006
16. Ertan, S., Lee, C., Willets, A., Tan, H., Pentland, A.: A wearable haptic navigation guidance system. In: Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215), pp. 164–165, Oct 1998
17. Fenza, G., Furno, D., Loia, V.: Enhanced healthcare environment by means of proactive context aware service discovery. In: 2011 IEEE International Conference on Advanced Information Networking and Applications, pp. 625–632. IEEE (2011)
18. Garlan, D., Siewiorek, D.P., Smailagic, A., Steenkiste, P.: Project Aura: toward distraction-free pervasive computing. IEEE Pervasive Comput. **1**(2), 22–31 (2002)
19. Giacomin, J.: What Is human centred design? ResearchGate **17**(4) (2014)
20. Golding, A.R., Lesh, N.: Indoor navigation using a diverse set of cheap, wearable sensors. In: Digest of Papers. Third International Symposium on Wearable Computers, pp. 29–36, Oct 1999
21. Goldstein, E.B., Brockmole, J.: Sensation and Perception. Cengage Learning (2016)
22. Gorbenko, A., Popov, V.: Anticipation in simple robot navigation and finding regularities. Appl. Math. Sci. **6**(132), 6577–6581 (2012)
23. Gorbenko, A., Popov, V.: The force law design of artificial physics optimization for robot anticipation of motion. Adv. Stud. Theor. Phys. **6**(13), 625–628 (2012)
24. Hawk, B., Rieder, D.M., Oviedo, O.O.: Small Tech: The Culture of Digital Tools. U of Minnesota Press (2008)
25. Hayden, D., Colbry, D., Black, J.A. Jr., Panchanathan, S.: Note-taker: enabling students who are legally blind to take notes in class. In: Proceedings of the 10th International ACM SIGACCESS Conference on Computers and Accessibility, Assets '08, pp. 81–88. ACM, New York, NY, USA (2008)
26. Hickey, J.V., Thompson, W.E.: I/M Society Focus. Addison-Wesley Educational Publishers (1995)
27. Hoffmann, H.: Perception through visuomotor anticipation in a mobile robot. Neural Netw. **20**(1), 22–33 (2007)
28. Hong, J., Suh, E.-H., Kim, J., Kim, S.: Context-aware system for proactive personalized service based on context history. Expert Syst. Appl. **36**(4), 7448–7457 (2009)
29. Hub, A., Diepstraten, J., Ertl, T.: Design and development of an indoor navigation and object identification system for the blind. In: Proceedings of the 6th International ACM SIGACCESS Conference on Computers and Accessibility, pp. 147–152. ACM, New York, NY, USA (2004)

30. Ishimaru, S., Kunze, K., Kise, K., Weppner, J., Dengel, A., Lukowicz, P., Bulling, A.: In the blink of an eye: combining head motion and eye blink frequency for activity recognition with google glass. In: Proceedings of the 5th Augmented Human International Conference, pp. 15:1–15:4. ACM, New York, NY, USA (2014)
31. Iwata, H.: Haptic interfaces. In: Jacko, J.A., Sears, A. (eds.) The human-computer interaction handbook, pp. 206–219. L. Erlbaum Associates Inc., Hillsdale, NJ, USA (2003)
32. Jaimes, A., Gatica-Perez, D., Sebe, N., Huang, T.S.: Human-centered computing: toward a human revolution. IEEE Comput. **40**(LIDIAP-ARTICLE-2007-001), 30–34 (2007)
33. Karray, F., Alemzadeh, M., Saleh, J.A., Arab, M.N.: Human-computer interaction: overview on state of the art (2008)
34. Kern, N., Schiele, B., Schmidt, A.: Multi-sensor activity context detection for wearable computing. In: Ambient Intelligence, pp. 220–232. Springer, Berlin, Heidelberg (2003)
35. Koesling, H., Kenny, A., Finke, A., Ritter, H., McLoone, S., Ward, T.: Towards intelligent user interfaces: anticipating actions in computer games. In: Proceedings of the 1st Conference on Novel Gaze-Controlled Applications, p. 4. ACM (2011)
36. Lane, N.D., Mohammod, M., Lin, M., Yang, X., Lu, H., Ali, S., Doryab, A., Berke, E., Choudhury, T., Campbell, A.: Bewell: a smartphone application to monitor, model and promote wellbeing. In: 5th International ICST Conference on Pervasive Computing Technologies for Healthcare, pp. 23–26 (2011)
37. Leidner, D.E., Kayworth, T.: Review: a review of culture in information systems research: toward a theory of information technology culture conflict. MIS Quart. **30**(2), 357–399 (2006)
38. Liu, J., Lockhart, T.E., Jones, M., Martin, T.: Local dynamic stability assessment of motion impaired elderly using electronic textile pants. IEEE Trans. Autom. Sci. Eng. **5**(4), 696–702 (2008)
39. Liu, Q.: Context-aware mobile recommendation system based on context history. Indonesian J Electr Eng Comput. Sci. **12**(4), 3158–3167 (2014)
40. Llaneras, R.E., Singer, J.P.: In-vehicle navigation systems: interface characteristics and industry trends. In: Proceedings of the Second International Driving Symposium on Human Factors in Driver Assessment, Training and Vehicle Design. University of Iowa, Iowa City, IA (2003)
41. Lu, H., Pan, W., Lane, N.D., Choudhury, T., Campbell, A.T.: SoundSense: scalable sound sensing for people-centric applications on mobile phones. In: ResearchGate, pp. 165–178 (2009)
42. Luczak, H., Roetting, M., Oehme, O.: The human-computer interaction handbook, pp. 187–205. L. Erlbaum Associates Inc., Hillsdale, NJ, USA (2003)
43. Mace, R.L.: Universal design in housing. Assistive Technol. **10**(1), 21–28 (1998)
44. Mann, S., Huang, J., Janzen, R., Lo, R., Rampersad, V., Chen, A., Doha, T.: Blind navigation with a wearable range camera and vibrotactile helmet. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 1325–1328. ACM, New York, NY, USA (2011)
45. Marion, A., Heinsen, E., Chin, R., Helmso, B.: Wrist instrument opens new dimension in personal information. Hewlett-Packard J. (1977)
46. May, A.J., Ross, T., Bayer, S.H., Tarkiainen, M.J.: Pedestrian navigation aids: information requirements and design implications. Personal Ubiquitous Comput. **7**(6), 331–338 (2003)
47. McDaniel, T., Krishna, S., Balasubramanian, V., Colbry, D., Panchanathan, S.: Using a haptic belt to convey non-verbal communication cues during social interactions to individuals who are blind. In: IEEE International Workshop on Haptic Audio visual Environments and Games, 2008 HAVE 2008, pp. 13–18. IEEE (2008)
48. McDaniel, T., Panchanathan, S.: The building blocks of somatic information delivery. In: Proceedings of the 19th ACM International Conference on Multimedia, pp. 877–878. ACM, New York, NY, USA (2011)
49. Michaelis, J.R., Rupp, M.A., Kozachuk, J., Ho, B., Zapata-Ocampo, D., McConnell, D.S., Smither, J.A.: Describing the user experience of wearable fitness technology through online product reviews. In: Proceedings of the 2016 annual meeting of the International Human Factors and Ergonomics Society (2016)

50. Miluzzo, E., Cornelius, C.T., Ramaswamy, A., Choudhury, T., Liu, Z., Campbell, A.T.: Darwin phones: the evolution of sensing and inference on mobile phones. In: Proceedings of the 8th International Conference on Mobile Systems, Applications, and Services, pp. 5–20. ACM, New York, NY, USA (2010)
51. Mistry, P., Maes, P.: SixthSense: a wearable gestural interface. In: ACM SIGGRAPH ASIA 2009 Sketches, pp. 11:1–11:1. ACM, New York, NY, USA (2009)
52. Muensterer, O.J., Lacher, M., Zoeller, C., Bronstein, M., Kübler, J.: Google glass in pediatric surgery: an exploratory study. Int. J. Surg. **12**(4), 281–289 (2014)
53. Muñoz, M.A., Gonzalez, V.M., Rodríguez, M., Favela, J.: Supporting context-aware collaboration in a hospital: an ethnographic informed design. In: International Conference on Collaboration and Technology, pp. 330–344. Springer (2003)
54. Nawyn, J., Intille, S.S., Larson, K.: Embedding behavior modification strategies into a consumer electronic device: a case study. In: UbiComp 2006: Ubiquitous Computing, pp. 297–314. Springer, Berlin, Heidelberg (2006)
55. Norman, D.A.: The psychology of everyday things. Basic Books (1988)
56. Norman, D.A., Draper, S.W.: User centered system design. New perspectives on human-computer interaction. L. Erlbaum Associates Inc., Hillsdale, NJ, USA (1986)
57. Oriolo, G., Ulivi, G., Vendittelli, M.: Real-time map building and navigation for autonomous robots in unknown environments. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) 28(3), 316–333 (1998)
58. Pantic, M., Pentland, A., Nijholt, A., Huang, T.S.: Human computing and machine understanding of human behavior: a survey. In: Artifical Intelligence for Human Computing, pp. 47–71. Springer (2007)
59. Parette, P., Scherer, M.: Assistive technology use and stigma. Educ. Train. Dev. Disabil. **39**(3), 217–226 (2004)
60. Patel, M.S., Asch, D.A., Volpp, K.G.: Wearable devices as facilitators, not drivers, of health behavior change. JAMA **313**(5), 459–460 (2015)
61. Pejovic, V., Musolesi, M.: Anticipatory mobile computing for behaviour change interventions. In: Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication, pp. 1025–1034. ACM, New York, NY, USA (2014)
62. Pejovic, V., Musolesi, M.: Anticipatory mobile computing: a survey of the state of the art and research challenges. ACM Comput. Surv. 47(3), 47:1–47:29 (2015)
63. Peli, E.: Visual issues in the use of a head-mounted monocular display. Optical Eng. **29**(8), 883–892 (1990)
64. Perera, C., Zaslavsky, A., Christen, P., Georgakopoulos, D.: Context aware computing for the internet of things: a survey. IEEE Commun. Surveys & Tutorials **16**(1), 414–454 (2014)
65. Poslad, S.: Ubiquitous Computing: Smart Devices, Environments and Interactions. Wiley (2011)
66. Prattichizzo, D., Chinello, F., Pacchierotti, C., Malvezzi, M.: Towards wearability in fingertip haptics: A 3-DoF wearable device for cutaneous force feedback. IEEE Trans. Haptics **6**(4), 506–516 (2013)
67. Rachuri, K.K., Musolesi, M., Mascolo, C., Rentfrow, P.J., Longworth, C., Aucinas, A.: EmotionSense: a mobile phones based adaptive platform for experimental social psychology research. In: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, pp. 281–290. ACM (2010)
68. Rett, J., Dias, J.: Human-robot interface with anticipatory characteristics based on Laban movement analysis and Bayesian models. In: 2007 IEEE 10th International Conference on Rehabilitation Robotics, pp. 257–268. IEEE (2007)
69. Rhodes, B.J.: The wearable remembrance agent: a system for augmented memory. Personal Technol. **1**(4), 218–224 (1997)
70. Rosen, R.: Anticipatory systems. In: Anticipatory Systems, pp. 313–370. Springer, New York (2012)
71. Rouse, M.: Nomadic computing (Mobile computing) (2007)

72. Sandell, G.J.: Review of auditory display: sonification, audification, and auditory interfaces. Music Percept.: An Interdisc. J. **13**(4), 583–591 (1996)
73. Stramler, J.H. Jr.: The dictionary for human factors/ergonomics. CRC Press (1992)
74. Swenson, H.N., Zelenka, R.E., Hardy, G.H., Dearing, M.G.: Simulation evaluation of a low-altitude helicopter flight guidance system adapted for a helmet-mounted display. In: IEEE/AIAA 10th Digital Avionics Systems Conference, pp. 115–124, Oct 1991
75. Tadayon, A., Zia, J., Anantuni, L., McDaniel, T., Krishnamurthi, N., Panchanathan, S.: A shoe mounted system for Parkinsonian Gait detection and real-time feedback. In: HCI International 2015 - Posters' Extended Abstracts, pp. 528–533. Springer, Cham, Aug 2015. https://doi.org/10.1007/978-3-319-21380-4_90
76. Thorp, E.O.: The invention of the first wearable computer. In: Digest of Papers. Second International Symposium on Wearable Computers (Cat. No.98EX215), pp. 4–8, Oct 1998
77. Warwick, K.: I, Cyborg. University of Illinois Press (2004)
78. Witt, H.: Context-awareness and adaptive user interfaces. In: User Interfaces for Wearable Computers, pp. 63–72. Vieweg+Teubner (2008)
79. Woods, R.L., Fetchenheuer, I., Vargas-Martin, F., Peli, E.: The impact of non-immersive HMDs on the visual field. SID Symp. Digest Techn. Papers **33**(1), 998–1001 (2002)
80. Yang, C.-C., Hsu, Y.-L.: A review of accelerometry-based wearable motion detectors for physical activity monitoring. Sensors **10**(8), 7772–7788 (2010)

Part V
Wireless Charging and Energy Transfer

Wireless Transfer of Energy Alongside Information in Wireless Sensor Networks

Hooman Javaheri and Guevara Noubir

Abstract Despite steady improvements over the last few decades, wireless communication networks are still severely constrained by the availability of an energy source. The problem manifests itself in many applications, in particular, wireless sensor networks enabling a plethora of Internet of Things devices, where communications are infrequent and the nodes are often idle; and also small-scale communication networks, whose nodes need to be minuscule limiting the possibilities to incorporate long-term energy peripherals such as batteries. Such applications can achieve optimal energy efficiency using passive (battery-less) receivers that wirelessly receive energy and information at the same time. In this chapter, we describe a set of techniques, introduced in our past work (Javaheri H Wireless Transfer of Energy Alongside Information: From Wireless Sensor Networks to Bio-Enabled Wireless Networks 2012 [37]), (Javaheri, Noubir, ipoint: A platform-independent passive information kiosk for cell phones 2010 [42]), to simultaneously deliver energy alongside information during wireless communications. We present mechanisms to consolidate energy and information transfer in wireless sensor networks. We introduce iPoint, a communication system including a passively powered wireless receiver capable of establishing two-way communication with a commodity smartphone without any hardware modifications. In contrast to traditional RFID tags, iPoint provides high computation and sensing capabilities and most importantly, it does not require specialized reader device to communicate. We prototype and experimentally evaluate our design that includes optimization techniques to ensure efficient delivery of energy and information and novel communication protocols.

H. Javaheri (✉) · G. Noubir
College of Computer and Information Science, Northeastern University,
Boston, MA 02115, USA
e-mail: hooman@ccs.neu.edu

G. Noubir
e-mail: noubir@ccs.neu.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_13

1 Introduction

Communication means imparting or exchanging of information among several parties. One important point, which is often overlooked, is that the parties engaged in any form of communication requires a certain amount of energy to send, receive, and process the traveling information. A dead node in a network never receives the message intended for it much like a dead person who will never hear a cry.

Over the last decade, wireless communication networks have achieved major success and emerged as the key technology for enabling the mobile revolution. Providing mobility and accessibility, wireless communication redefines the notion of network and connectivity, and powers a huge range of applications. The speed, capacity and robustness of wireless communication keep improving every day, but several challenges, most notably energy efficiency, hinder their effectiveness [22, 46]. Conserving energy in wireless networks is particularly important because the wireless nodes are critically dependent on their limited energy resource, that is, their battery. Receiving and decoding a wireless signal normally requires a significant amount of computations. In addition, transmitting wireless messages is one of the most energy consuming tasks in today's computing. Performing such energy-hungry tasks with low efficiency reduces the lifetime of the nodes and eventually causes node death degrading connectivity and performance of the network. These challenges are specially critical for the emerging IoT revolution powered by wireless sensor networks. Therefore, several attempts at hardware and software levels have been made to tackle this problem: more efficient batteries with larger capacities have been invented [54]; the power consumption of wireless nodes has been significantly reduced thanks to advances in low-power electronics and new methods of computation such as reversible computing [6]; finally, many energy-aware communication protocols and algorithms that factor in the energy constraints of wireless nodes have emerged [17, 74, 80, 81].

All wireless networks struggle with the energy conservation issue, yet the problem is more evident in certain applications such as wireless sensor networks (WSNs). A WSN usually consists of several spatially distributed nodes that monitor or sense a physical or environmental condition, and transmit the collected data via wireless communication. The sensing process generally does not require a lot of energy. In fact, wireless transmission accounts for almost all the node's energy consumption. In most cases, the data collection occurs sporadically or upon an external request. Therefore, continuous operation of a node's radio, which results in fast battery drainage, is not necessary. Current techniques rely on periodically waking the receiver up to synchronize and respond to the requests of a master node [13, 91]. These methods help conserve energy but are far from optimal. Ideally, the radio component of the sensor node should go into a full-sleep (idle) mode that consumes virtually no energy and wakes up only on external requests or events. This can be achieved by integrating

a completely passive component that acts as a wake-up circuit and relays the external request to the idle node. The challenge of engineering such a system, which includes software, hardware, and communication protocols, is one of the goals of this study.

Having a receiver that consumes *no energy* while being idle is the optimal solution for any of the scenarios mentioned above. Note that, this goal can not be achieved by simply reducing the energy consumption of the radio components; rather, the energy consumption should be eliminated when no communication is taking place. This rules out the use of conventional wireless devices which are basically an ensemble of active electronic boards that consume energy while operating, regardless of how energy-efficient they are. Instead, let us consider the following scenario: assume that the incoming signal, which carries a certain amount of information, provides the receiver with the energy required to extract the information. On the other end, the receiver simultaneously executes two actions on the signal. First, it converts the energy of the signal into a usable form. Second, it runs the decoding procedure to extract the embedded information. In this case, the energy consumption occurs only when there is an incoming signal, and it is fully provided by the transmitter entity. Therefore, the receiver does not rely on any other source of energy.

Combining energy and information transfer is a promising approach that leads to engineering *passive* wireless receivers. There are many communication schemes that provides a vast range of throughput, complexity, and energy efficiency. Also, a few mechanisms such as RF energy harvesting have been proposed to transfer energy wirelessly [53, 86]. However, combining these schemes presents several challenges. The range and capacity of today's energy transfer methods are very limited, which makes pairing them with normal communication methods difficult and often impossible. In order to have a functional system, we need to carefully optimize, modify, or completely revamp the energy and information transfer mechanisms. This includes building more efficient systems using specialized hardware, and devising better algorithms software solutions that exhaust the physical limits of the hardware.

In this chapter, we aim to explore several techniques to transfer energy alongside information via a wireless link. We look into techniques to consolidate the energy and information transfer in wireless networks. We review and compare the energy transfer technologies, provide design considerations, and sketch guidelines to implement such functionality. At the end, we present a communication system that allows two-way communication between a commodity smartphone and a passive receiver. The system features a combination of ultra-low-power electronics and RF energy harvesting. We introduce several techniques to increase the efficiency of the information and energy channels, propose novel communication paradigms and protocols optimized for our setup, build prototypes and finally, evaluate the system performance experimentally.

The contents of this chapter is based on authors' previous work [37, 42]. The results, discussions, and figures may have appeared in other manuscripts published by the authors.

Chapter Organization The rest of this chapter is structured as follows:
In Sect. 3, we define a model for consolidated energy and information channels in wireless sensor network. We describe the framework for our model and review the

building blocks of such model along with its design requirements. This section serves as the basis for Sect. 4 in which we present iPoint.

Section 4 includes a comprehensive design, optimization, prototype, and performance evaluation of iPoint, a novel communication system that features a consolidated energy and information channel. This includes the design and characterization of hardware, software, and communication paradigms. Several optimization techniques including theoretical analysis and experimental validations are detailed. A shorter version of the results presented in this section has been published in [42].

Section 2 reviews the related literature. We discuss previous studies in the field, which presents the state-of-the-art for relevant technologies and identify the similarities and distinctions of our work. Section 5 concludes the document and presents the direction for future work.

2 Related Work

2.1 Radio Frequency Identification (RFID)

RFID tags have the potential to deliver information anytime, anywhere [23, 88]. However, RFID tags have significant limitations making them impractical for delivering a substantial amount of information to commodity smartphones. First of all, equipping smartphones with an RFID reader is a significant and challenging modification to the phone hardware. Second, among the three types of RFIDs (i.e., passive, active, and semi-active), only the passive ones do not require a battery and therefore satisfy severe longevity constraints. However, passive RFIDs require the readers to transmit at high power (in the order of watts), with large antennas. Furthermore, such RFIDs are only capable of storing a very limited amount of information (e.g., 128 bytes) and are not capable of sophisticated interactions.

2.2 WISP

Several RF energy harvesting techniques and prototypes were explored over the last few years. The WISP platform and its variants harvest energy from RFID reader [83] and TV radio stations [9], and are capable of powering an ultra-low-power microcontroller. The WISP was also used as a battery-less sensor node to communicate with a traditional RFID reader [83]. It relies on a high energy source (30 dBm) operating at a medium RF frequency (915 MHz). The constraint of the iPoint to operate on the low RF energy from smartphones (few dBm) and higher Wi-Fi frequency (i.e., 2.4 GHz) requires more advanced RF energy harvesting mechanisms that we present in the next sections. Other platforms for wireless power transfer exist but require either high transmission power on the 915 MHz band [69], or require highly customized

transmitters and receivers such as in wireless power transfer via strongly coupled magnetic resonances [50].

2.3 Backscattering

Communication based on RF backscattering has been an active avenue of research with application in medical sciences. In this form of communication, the transmitter sends RF signals to a passive device, normally implanted in the body. A receiver will measure the backscattered signal coming from the implant. With a careful design, one can modulate the information on the device in the backscattered signal. Researchers have shown a proof-of-concept design that can convert wireless transmissions from one technology (Bluetooth) to another (Wi-Fi, Zig-Bee) using backscatter communication via an implanted device [34]. Such design enables multimodal communication between implanted devices such as contact lenses and wearable devices equipped with multiple RF technologies in order to report medical or diagnostic information from the body.

2.4 Multi-path Energy Routing

Recent research [61] has shown promising results in transferring RF energy via a network of energy-harvester nodes. Energy can be routed alongside data in such networks. Within a network consisting of energy harvesters that are optimized to accumulate and relay the RF energy in different frequency bands, the energy flows along multiple paths increasing the overall harvesting efficiency.

2.5 Near-Field Communication (NFC)

The near-field communication (NFC) [62], founded by Nokia, Phillips, and Sony in 2004, is a set of standards based on RFID technology that allows devices such as smartphones to establish a communication. Communications take place at 13.56 MHz over very short range of less than few centimeters, which typically involves two devices touching each other. NFC allows several modes of communications including communication with a passive device. The communication scheme is ASK with Miller or Manchester coding. Data rate varies between 106 and 424 kbit/s. Because of the very simple and fast set-up, NFC can be used to initialize more sophisticated wireless communication to start peer-to-peer networking. The energy transfer mechanism is based on electrodynamic induction between two loop antennas. In comparison to the solution presented in this work, NFC exchanges data at higher data rates but operates at shorter ranges. Devices communicating over NFC normally needs to touch to

ensure reliability. Moreover, NFC functionality on a smartphone requires additional designated hardware, which adds another hardware dependency.

2.6 Bokode

Recently, a clever alternative solution to RFID tags and traditional barcodes, called bokode, was developed to deliver information from a dot of 3 mm diameter encapsulating a high-density Data Matrix code [60]. The information is revealed by putting an off-the-shelf camera in an out of focus mode. This solution has the advantage to reduce the size of the tag and increase the information density but still keeping the tag passive. However, bokode still lacks a two-way communication capability and requires sophisticated digital cameras (ten megapixel with a large lens) with in/out focus capability. In the future, if smartphones become equipped with controllable focus cameras, the envisioned iPoint system might benefit from integrating a bokode-based LCD display to deliver information to a smartphone at a lower energy cost.

2.7 Microwave Power Transmission

A mechanism based on RF energy harvesting has been explored to transfer a huge amount of energy over very long distances [58, 93]. For example, high-power directional antennas have been used to transmit energy to satellites from earth. The rectennas with efficiencies up to 95% have been demonstrated [85].

2.8 Resonant Inductive Coupling

Nonresonant inductive coupling, which is being widely used in transformers, suffers from stiff drop in efficiency when the distance between interacting coils increases. Karalis et al. [50] have shown that extremely high efficiency over mid-range can be achieved if the inductive coils couple in their resonant frequency. They have built a system that can transfer 60 W of power over distances up to 2 meters with 90% efficiency. Their work is the basis for a commercial technology called WiTricity [4] that promises efficient wireless power transfer that can be used to wirelessly power electronic devices within environments such as home or office. Several methods based on resonant inductive coupling have been proposed for wireless energy transfer in the biological setting, for instance to power medical implants in body [68, 73].

3 Consolidated Energy and Information Channels (CEICh) in Wireless Sensor Networks

In the previous section, we argued that combining energy and information transfer may provide an optimal solution to energy conservation in wireless sensor networks. In this section, we overview techniques to combine energy and information transfer channels in wireless sensor networks.

3.1 Overview of CEICh

First, we define the framework (model) in which CEICh is being implemented as follows:

- We consider a wireless sensor network whose nodes (devices) perform computations electronically and communicate using radio frequency (RF) signals.
- The network includes two types of nodes: passively powered receivers and the energy-provider master node, which we call the source.
- The communication only occurs between a source and a receiver at the time. We do not consider the communication among receiver nodes.
- In order to eliminate the energy conservation problem, the receiver may not rely on a battery or other unpredictable source of energy, such as solar energy or mechanical vibrations.
- In one communication cycle, the source sends the request to the receiver; the receiver receives the request, processes it, computes the reply, and sends it back to the source.
- The signal from source to receiver contains both energy and information. The receiver obtains the energy required to receive and decode the transmitted signal from the signal itself. Signals transmitted by receiver are not required to contain energy.

Figure 1 illustrates a general schematic of a CEICh.

3.2 Design of CEICh

A CEICh-enabled system performs three tasks: energy transfer, communication, and computation. In this section, we take a closer look into the important features and design elements of a CEICh enabled system.

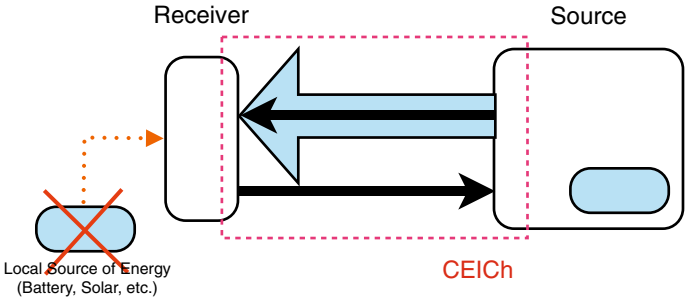


Fig. 1 The general schematic of CEICh

3.2.1 Energy Transfer Mechanisms

Mechanisms to transfer energy from a power source to a wireless device fall into two categories:

Electrodynamic Induction Methods based on electrodynamic induction use inductive coupling between the source and the receiver, and provide high-efficiency energy transfer over very short distances. In such systems, the efficient energy transfer occurs in the near-field region of the source's inductive antenna. The near-field region boundaries are determined by the size of the antenna and the wavelength of the transmission. For smaller antennas, the distances less than a wavelength, $R < \lambda$, from the source is considered near-field region, while the transition to far-field occurs in $\lambda < R < 2\lambda$. For larger antennas, the near-field region boundary is expressed by Fraunhofer distance:

$$R < \frac{2D^2}{\lambda}, \tag{1}$$

where D is the largest dimension of the antenna. The relationship between transmit power and the distance in the near field can be expressed as follows.

$$P_{\mathrm{T}} \propto \begin{cases} R^{-1} \text{ for } & R \leq 0.1\lambda \\ R^{-3} \text{ for } & R > 0.1\lambda \end{cases} \tag{2}$$

The efficiency declines significantly over greater distances, which makes such methods ineffective when the signal has high frequency and long range, a typical case in communications. Recently developed methods apply resonant inductive coupling to achieve much higher efficiency over longer ranges.

Electromagnetic Radiation Methods based on electromagnetic radiation, also called RF energy harvesting, include receivers equipped with *rectenna*, a specialized component made of an antenna connected to a voltage rectifier that converts the electromagnetic energy of the received signal to a usable DC voltage. These methods operate in far-field region of the corresponding antennas hence attain longer ranges. However, they exhibit a number of limitations. At the antenna level, a system

using omnidirectional antennas shows a quadratic drop in efficiency with respect to the distance. Moreover, the losses due to imperfect rectification and antenna matching reduces the efficiency of the system. Matching and rectification losses can be minimized in a well-designed rectenna. Also, the antenna efficiency improves using directional antennas and applying beam-forming techniques.

3.2.2 Energy Transfer Efficiency

The most important property of any energy transfer mechanism is the efficiency, that is what portion of the received energy from the source is converted to usable energy for the receiver side. The factors determining the overall efficiency of an energy transfer system is different depending on the method of transfer.

In case of electrodynamic induction, the overall efficiency is chiefly determined by mutual inductance between the interacting coils, which in general is given by the Neumann formula:

$$M_{ij} = \frac{\mu_0}{4\pi} \oint_{C_j} \oint_{C_i} \frac{\mathrm{d}s_i \cdot \mathrm{d}s_j}{|R_{ij}|} \tag{3}$$

where M_{ij} is the mutual inductance between coil i and coil j, μ_0 is the vacuum permeability, C_i and C_j denote the curve of the coils and R_{ij} is the distance between two points. The final value of the mutual inductance depends on the size of the resonators, their relative orientation, and the distance between them. The coupling coefficient between inductive resonators is defined as:

$$k = \frac{M_{ij}}{\sqrt{L_i L_j}}, \tag{4}$$

where L_i and L_j are self-inductance of the resonators. In many common inductive coupling systems, $0 < k < 1$ and the system is called weakly coupled. Using recent methods such as *Resonant Inductive Coupling*, one can achieve higher coupling factors ($k > 1$) and build strongly coupled inductive resonators. Other contributing factors in overall efficiency includes ohmic and core losses in the coils, and appropriate impedance matching at either side.

In energy harvesting systems, the overall energy transfer efficiency can be estimated by aggregating the effect of signal's transmission losses, impedance mismatch, and rectifier's energy efficiency. Signal transmission losses are given by Friis equation, which gives P_{R}, the power received at receiver's antenna:

$$P_{\mathrm{R}} = P_{\mathrm{T}} G_{\mathrm{t}} G_{\mathrm{r}} L_{\mathrm{p}} \left(\frac{\lambda}{4\pi R} \right)^2, \tag{5}$$

where P_{T} is transmit power; G_{t} and G_{r} are antenna gains at transmitter and receiver side, respectively; L_{p} is the polarization loss; λ is the wavelength, and R is the

distance between the antennas. Adding the effect of impedance mismatch, we can calculate the power that enters the rectifier circuit, P_{rec} as:

$$P_{\text{rec}} = (1 - |\Gamma_r|^2) P_{\text{R}}, \tag{6}$$

where Γ_{r} is the reflection coefficient of the receiving antenna. Finally, the input power to the receiver is given by:

$$P_{\text{in}} = \eta_{rec} \times P_{rec}, \tag{7}$$

where η_{rec} is the efficiency of the rectifying circuit. The value of η_{rec} depends on the design of the rectifier (e.g., half-wave or full-wave) as well as the electrical characteristics of its components such as forward and break voltages of the diodes and leakage voltage of the charging capacitors. In practice, the efficiency of the rectifier is measured experimentally since the exact calculation proves impractical due to complexity and nonlinearity of the circuit. In summary, the overall efficiency can be expressed as the following:

$$\eta = \frac{P_{\text{in}}}{P_{\text{T}}} = \eta_{\text{rec}} G_{\text{t}} G_{\text{r}} L_{\text{p}} \left(\frac{\lambda}{4\pi R} \right)^2 (1 - |\Gamma_{\text{r}}|^2). \tag{8}$$

Each of the contributing elements mentioned in Eq. 8 may be optimized to achieve higher energy transfer efficiency in the design of a CEICh. It is worth mentioning that Eq. 8 holds true assuming the perfect channel conditions. In practice, the efficiency is further affected by fading (F^{-1}), shadowing (B^{-1}), and on-object antenna gain penalties (Θ^{-1}), each of which can be either modeled or measured.

3.2.3 Signal Amplification

Another important factor in the design of a CEICh is the minimum voltage required to power-up the receiver, V_{min}. If the output voltage of the energy transfer unit is not sufficient, a voltage multiplier circuit may be used to elevate the voltage beyond V_{min}. Because of the limited energy budget, the voltage multiplication is done using a passive circuit. In energy harvesting mechanisms, signal rectification and amplification is done using voltage multipliers.

3.2.4 Voltage Rectification

A voltage rectifier, which is typically a network of diodes and capacitors, rectifies an input AC signal to output DC voltage. The simplest rectifying circuit is the envelope detector circuit shown in Fig. 2. Using a diode with forward voltage of V_{D}, the DC voltage of $V_{\text{out}} = |V_{\text{in}}| - V_{\text{D}}$ can be obtained in no-load conditions. The circuit

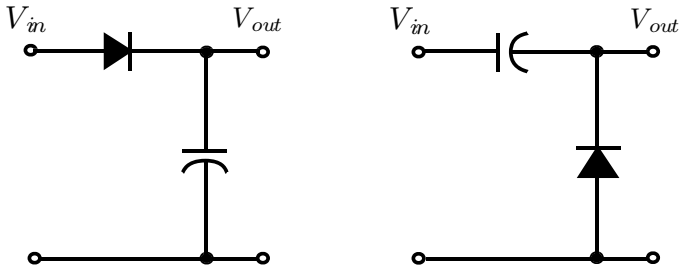


Fig. 2 The schematics of envelope detector circuit (left) and the clamp circuit (right)

experiences voltage ripples if connected to a load. This circuit is also called half-wave rectifier because only the positive half of the AC waveform is rectified.

A clamp circuit as shown in Fig. 2 may be used to achieve higher DC voltage levels. The negative half of the waveform is clamped to zero, and the output voltage of $V_{\text{out}} = 2|V_{\text{in}}| - V_{\text{D}}$ is obtained. The circuit, however, shows a very drastic ripple of size $2V_{\text{in}}$ when connected to a load.

By placing a clamp circuit in series with an envelope detector, one can achieve higher DC voltage levels and significantly improves the ripple characteristics of the circuit. This circuit, shown in Fig. 3, is called voltage doubler and was first invented by Heinrich Greinacher in 1913, and is later used to power the particle accelerators [16]. Let us look at the steady-state analysis of the voltage doubler circuit. On the negative half of the input AC waveform, the capacitor C_1 is charged until $V_{\text{C1}} = |V_{\text{in}}| - V_{\text{D}}$. Similarly, over the positive half, the capacitor C_2 is charged until we have

$$V_{\text{out}} = V_{\text{C2}} = |V_{\text{in}}| - V_{\text{D}} + V_{\text{C2}} \tag{9}$$

$$= 2(|V_{\text{in}}| - V_{\text{D}}) \tag{10}$$

The key feature of the Greinacher circuit is that individual voltage doublers can be cascaded to build a voltage multiplier that generates arbitrarily large output DC voltage. For an N-stage voltage multiplier, the output voltage of $V_{\text{out}} = 2N(|V_{\text{in}}| - V_{\text{D}})$ may be obtained theoretically. A output voltage can also be doubled by placing

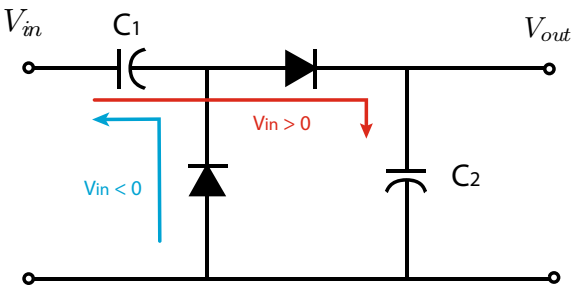


Fig. 3 Schematic of Greinacher voltage doubler

the voltage doubler in parallel with a mirrored copy with the same topology but reversed polarity of the diodes. In a careful design, the amount of ripple can be minimized when connecting the multiplier to a load. Note that, the analysis above assumes that the diodes and capacitors in the circuit are ideal in no-load conditions. In nonideal scenario, the parasitic effect of the capacitors in each stage and nonlinearity of the diodes significantly affects the performance of the voltage multiplier. Also, the behavior of the components will depend on the output current (load). There is no closed form for the efficiency of the system in realistic conditions. Therefore, the optimization of the overall efficiency of the rectifier should be based on simulation and experimental measurement.

3.2.5 Signal Characteristics

Incorporating energy transfer in communications requires modification in signal characteristics. Unlike conventional communication systems, the receiver does not include an active signal amplifier, therefore the transmitter has to provide a signal with higher level of energy to maintain the receiver in operating zone. In contrast, response signals do not have the aforementioned restriction since the master node does not have any energy constraints and may possess an active radio receiver with signal amplifier to receive and demodulate the response from the receiver.

Signal Preambles, Trailers, and Energy Storage In most cases, the converted power obtained from the signal is not sufficient for continuous operation of the receiver, hence the converted energy should be stored and accumulated over a longer period. In particular, the receiver needs to obtain a minimum amount of energy before it can demodulate the information. Therefore, the signal should include a *preamble* designed to provide the energy to start-up the receiver. Similarly, having received the signal, the receiver continues to process and respond to the request. The receiver should have already harvested enough energy to complete the tasks, otherwise, the signal needs to include a *trailer* that provides the required energy and may contain no actual information. We can define a duty cycle for the system which can be estimated by the following ratio:

$$DC = \frac{P_{\mathrm{ET}}}{P_{\mathrm{active}}} = \frac{\eta P_{\mathrm{T}}}{P_{\mathrm{active}}}, \tag{11}$$

where η is the energy transfer efficiency, P_{T} denotes the power sent by the transmitter, P_{ET} is the converted power from energy transfer unit, and P_{active} is the average power that the receiver requires to perform a communication cycle, that is, to start-up, demodulate, process, and reply to the signal. This unavoidable overhead significantly reduces the data rate and throughput of the communications in a CEICh.

3.2.6 Communications Schemes

Because of severe energy constraints, the demodulation process has to be simple as possible, while remaining effective. Sophisticated decoding mechanisms demand a significant amount of computations and energy, hence are not optimal. Due to integration of the energy transfer and communication, the signal from the master node to the receiver carries a considerable amount of energy. Therefore, the signal-to-noise ratio (SNR) is quite high, which allows using simpler modulation and demodulation schemes.

Most of today's communication protocols do not fulfill the requirements mentioned above, thus it needs to be modified or completely revamped according to the design requirement and constraints. For example, frequency-shift keying (FSK) and phase-shift keying (PSK) mechanisms require fairly complex receiver design, hence are not suitable for communication in CEICh. On the other hand, schemes such as amplitude-shift keying (ASK) or on-off keying (OOK) can be detected by a simpler receiver design (an envelope detector circuit), and consequently can be used in CEICh with minor modifications. Backscattering techniques are used to send information from the receiver to the source. Such techniques based on modulating information in the amplitude of reflected waves adjusting the impedance of the receiver's antenna, allow highly efficient bidirectional communication without using a stand-alone RF transmitter in the receiver. In the following section, we present a system that employs a two-way consolidated energy and information channel. The system includes two optimized communication protocols to achieve maximum energy efficiency.

4 IPoint

In this section, as an implementation of a CEICh-enabled wireless system, we propose a system that provides two-way communication between a receiver with no source of energy and a conventional smartphone. We present a complete design of the system featuring energy harvesting as the energy transfer method. We explore various technologies, introduce new communication paradigms, and build a prototype of the passively powered device that we call *iPoint*.

4.1 Motivation and Possible Applications

Providing information anytime, anywhere, and to anybody is a challenging task. Consider an application where we want to deliver information to any person equipped with a standard smartphone in even remote locations where there is no network coverage, and where no source of energy is available. The information should be

delivered from a device that can last decades without maintenance. Examples of applications include information delivery to hikers lost in the woods (e.g., directions, closest points for assistance), caves, and also high-density and interactive information tags (e.g., tourists information, museums).

A smartphone is an ultimate example of a wireless device. It provides a remarkable combination of data acquisition, computing power, and communication interfaces, all in a highly portable package. Our design exploits such capabilities of the smartphone to mitigate energy conservation issues.

4.2 Definition of the System with Respect to CEICh Design

iPoint is a passively powered device that is capable of communicating with a commodity smartphone while obtaining the required energy for computation and communication from the signal. This completely fits the definition of the receiver in our CEICh model. Here, the master node or the source is a smartphone. In the process of designing the iPoint, in addition to considerations stated in Sect. 3, we focused on the following key characteristics:

- *Universality*: any smartphone should be able to serve as the master node without any hardware modifications. Installing a software application should be sufficient to enable all desired functionalities.
- *Interactivity*: the device should be able to accept, process, and reply to specific requests. In other words, the communication between the iPoint and the smartphone is bidirectional.

4.3 Challenges and Approach

The key challenge to implement an efficient CEICh channel is sever power deficiency throughout the system. Smartphones, unlike RFID readers, are not capable of transmitting high-power RF signals. In fact, the amount of power transmitted by cell phones is aggressively controlled because of several FCC regulations health issues. Another obstacle that emerges in the course of the design, is lack of specialized wireless interfaces in the smartphone. To maintain the universality of the system, the CEICh design should use one of the standard wireless interfaces of the smartphone (GSM, Wi-Fi, and Bluetooth) without any hardware modifications. This eliminates the use of several mechanisms described in Sect. 3. For example, inductive coupling energy transfer may not be implemented since the majority of smartphones do not include inductive coupling resonator antenna.

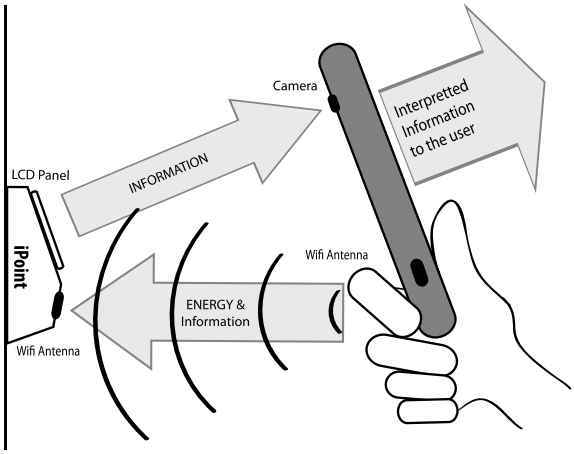


Fig. 4 Conceptual illustration of how iPoint device performs

4.4 Our Solution

These defining features lead us to a design that introduces innovative communication paradigms and techniques and the integration of a set of fairly unrelated technologies. Among the wireless interfaces present in smartphone, we picked Wi-Fi interface because it provides a greater degree of control by software, works on an unlicensed frequency band (2.4 GHz) and also has the highest transmit power. A backscattering technique cannot be used because of limited power budget and lack of compatible wireless interface to detect the backscattered signal in the smartphone. Instead, we design the iPoint to display the result on an LCD panel, exploiting the camera on the smartphone to acquire the data via the captured image of the display and interpret the information. Figure 4 shows how this design operates. In the following, we briefly review the system hardware and software architecture, the components, and the communication paradigms and techniques:

- *iPoint components*: the iPoint consists of a rectenna optimized for the 2.4 GHz Wi-Fi band, an ultra-low-power microcontroller with an LCD driver, and a multi-segment LCD panel.
- *Smartphone*: virtually any smartphone with a Wi-Fi network interface and an integrated camera.
- *Energy-provisioning*: the smartphone delivers the energy to the iPoint via Wi-Fi transmission. The iPoint benefits from a more efficient RF energy harvesting circuit optimized for limited transmission power of smartphones, about two orders of magnitude less than conventional RFID readers.
- *Multimodal communication*: we propose two novel communication mechanisms to circumvent the severe energy asymmetry and constraints. (1) The information from the smartphone to the iPoint is encoded in the Wi-Fi packet width, which results in much simpler and more energy-efficient demodulation at the expense of

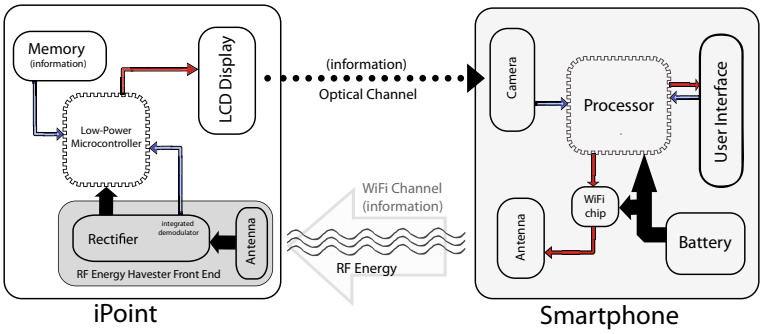


Fig. 5 Detailed diagram of iPoint

a lower data rate. (2) The information from the iPoint to the smartphone is encoded as a series of patterns shown on the LCD, to be captured by the smartphone camera.

4.5 Detailed System Architecture

This section outlines the architecture of the proposed system. We describe the system components in detail, discuss the required features, important parameters and trade-offs, and compare design choices. We break down hardware of the iPoint into three components: a rectenna that receives the information and energy, an ultra-low-power computing core to process the data, and a display to show the outputs (Fig. 5).

4.5.1 Energy Transfer

The energy transfer mechanism in iPhone is based on electromagnetic energy harvesting, therefore consists of a carefully design rectenna. The rectenna serves as the power supply for the device. Its two main components, the antenna and the rectifier circuit, in addition to two auxiliary circuits are described below:

- *Antenna*: The antenna is designed for the 2.4 GHz band. External whip antennas provide larger gain and better performance, whereas the integrated printed antenna make more compact design possible. The directionality of the antenna may be adjusted to achieve larger gain given spatial coordinates of the antenna with respect to the smartphone.
- *Rectifier circuit*: It converts the RF energy of the arrived Wi-Fi signals to DC voltage by passing them through a cascade of voltage multiplier circuits. Input power of the rectifier is often extremely small, therefore a multi-stage rectifier is used to build up sufficient output DC voltage, normally 1–5 V, to power the computing unit. The efficiency of the rectifier depends on the overall design of the circuit as

well as electrical characteristics of its components, notably forward voltage of the diodes and leakage of the capacitors. A full-wave rectifier shows better converting efficiency and produces more stable DC output compared to a half-wave rectifier, but requires a differential output design. Designs using Schottky diodes, which have smaller forward voltage, and RF optimized capacitors show significantly better performance.

- *Matching circuit*: The antenna and the rectifier should be carefully matched over the Wi-Fi frequency band. Matching is typically done by experiment. The trade-off between a good match and bandwidth of the rectenna should be considered in the design. The ideal bandwidth of the system is roughly 20 MHz, equal to width of the Wi-Fi channel.
- *Regulatory Circuit*: A shunt voltage regulator is placed after the rectifier is used to maintain the output DC voltage level within the safe range of operation for the computing unit.

Additionally, the rectenna constructs the envelope waveform of the arriving Wi-Fi transmission and passes it, as partially demodulated data, to the computing unit for further processing. We will discuss this in greater detail in Sect. 4.6.

4.5.2 Computation

The computing unit of the iPoint should provide extremely low-power consumption along with moderate computing capacity in a simple hardware design. Therefore, ultra-low-power microcontrollers (MCU), such as TI MSP430 family, are favorable design choices. The MCU should provide an adequate I/O interface and preferably include integrated drivers for external displays. Considering the energy constraints, the MCU may be underclocked to further reduce the power consumption.

4.5.3 Display

The iPoint displays the information with sufficient contrast and clarity to guarantee error-free pattern recognition by the smartphone, and minimize the energy consumption of the process. Passive-matrix liquid crystal displays (LCDs) [84] require a very small amount of energy to reach desirable contrasts without the need for a light source on the device, therefore is a better choice compared to LEDs. Given the same input voltage and distance from the camera, larger panels produce more pixels but less contrast compared to smaller panels.

4.6 Multimodal Communications

Because of the very low-power transmission of smartphones, the energy harvested by the rectenna is not sufficient to power a wireless transceiver and use conventional communication schemes. In this section, we describe two novel schemes that allow two-way communications between the iPoint and the smartphone within such limited energy levels. These schemes leverage smartphone capabilities to reduce the energy budget of the communications. We break down the communications into two separate channels: smartphone-to-iPoint (S2I) and iPoint-to-smartphone (I2S). We propose a different communication scheme for each channel.

4.6.1 Packet Length Modulation (PLM)

To provide energy for the iPoint device, the smartphone transmits Wi-Fi signals. However, demodulating Wi-Fi packets needs an active demodulator requiring energy beyond what iPoint harvester can obtain from the signals. We propose packet length modulation (PLM), a scheme in which the information is embedded in the length of the Wi-Fi packets.

Encoding In PLM, each packet length represents a symbol in the code. A message is defined as a sequence of Wi-Fi packets each representing its corresponding symbol. Note that, the smartphone sends the packets via its Wi-Fi interface, which uses the Wi-Fi protocol and does not know about the PLM. The following modifications to existing Wi-Fi protocol are necessary to implement PLM encoding functionality:

- Since a Wi-Fi access point may not be available, the smartphone creates an ad hoc Wi-Fi network.
- The iPoint lacks any Wi-Fi transceiver, hence no acknowledgment packet is sent back to the smartphone. Therefore, PLM uses broadcast packets to prevent unnecessary packet retransmission.
- The Wi-Fi packet length depends on the size of the packet, rate of the communication, and fragmentation. To have a robust encoding, the Wi-Fi interface should transmit at a fixed rate without using any rate-adaptation algorithm. This can be achieved by a UDP broadcast transmission.
- Assuming the PLM mechanism uses M different packet lengths, each packet encodes $\log_2 M$ bits of information. The fragmentation threshold determines the maximum packet length, hence the number of the symbols, and should be set to the maximum.

All modifications above are made in the software; no hardware modification is necessary.

Decoding To decode the PLM signal, the iPoint retrieves the length of the received Wi-Fi packets. First, the rectenna generates the envelope signal of the received Wi-Fi packet, and sends it to the computing unit. The computing unit samples the signal,

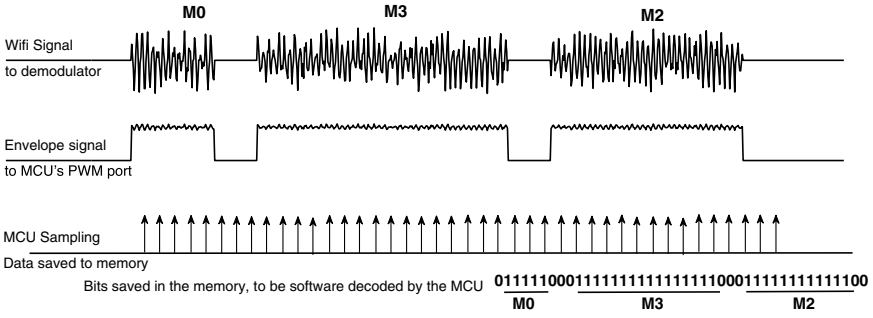


Fig. 6 PLM decoding

determines its length, and maps it to its corresponding symbol. Predefined start and stop flags can be used to distinguish the beginning and the end of a message. Figure 6 illustrates the steps of PLM decoding.

Data Rate Analysis Let M denote the number of the packet lengths in the PLM. Assume the transmitter sends the packets at rate, R. Let $S_{\min}$ be the smallest packet size. $S_{\min}$ is determined by the MCU clock to guarantee packet detection and length estimation, and by the energy-harvester efficiency and MCU energy requirements. We consider packets of size multiples of $S_{\min}$, $S_{\mathrm{i}} = i \times S_{\mathrm{m}in}$. Hence, the average size of the packet would be

$$S_{\mathrm{avg}} = \frac{S_{\min} + S_{\max}}{2} = \frac{(M+1)}{2} S_{\min}. \tag{12}$$

Further assume a message is a sequence of packets separated by *idle* periods of the length S_{idle}. We can calculate the time needed to send a packet, T_{p} and are as follows:

$$\begin{aligned} T_{\mathrm{p}} &= \frac{S_{\mathrm{avg}} + S_{\mathrm{idle}}}{R} \\ &= \frac{(M+1)S_{\min} + 2S_{\mathrm{idle}}}{2R} \end{aligned} \tag{13}$$

where R is the data rate of the Wi-Fi transmission in bps. Each packet encodes $\log_2 M$ bits of information. Therefore,

$$R_{\mathrm{PLM}} = \frac{\log_2 M}{T_{\mathrm{p}}} = \frac{2\log_2 M \times R}{(M+1)S_{\min} + 2S_{\mathrm{idle}}} \tag{14}$$

The values of $S_{\min}$ and S_{idle} are determined by the maximum sampling rate of the MCU at iPoint side. Assuming f_{MCU} is the sampling frequency of the MCU, we have

$$S_{\min}, S_{\mathrm{idle}} > \frac{2R}{f_{\mathrm{MCU}}} \qquad \text{(Nyquist theorem)}. \tag{15}$$

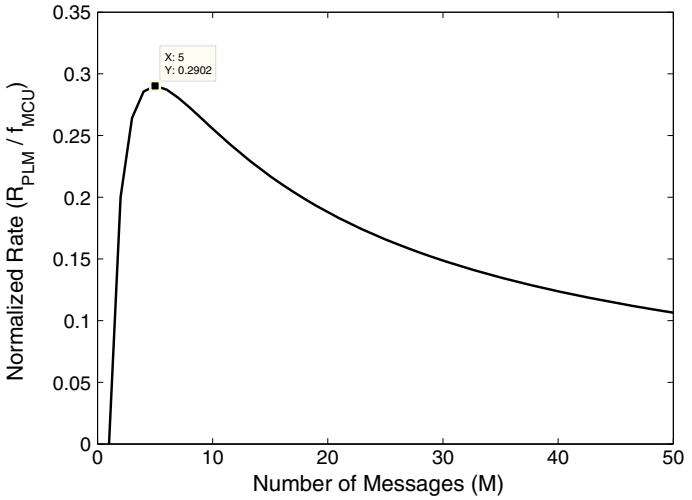


Fig. 7 Normalized Rate of PLM for different number of messages. It is shown that the best performance is achieved when $M = 5$

Hence,

$$R_{\text{PLM}} < \frac{2R \log_2 M}{(M+1)(\frac{2R}{f_{\text{MCU}}}) + 2(\frac{2R}{f_{\text{MCU}}})} \tag{16}$$

$$R_{\text{PLM}} < \frac{\log_2 M \times f_{\text{MCU}}}{M+3}. \tag{17}$$

Figure 7 illustrates the performance of the PLM for different values of M, showing that the best performance is achieved using five different packet lengths. Note that during idle periods, iPoint does not receive energy from the smartphone. This lowers the output voltage of the rectifier which may result in unwanted shutdown of the system. To maintain the harvested voltage level above the desired threshold, S_{min} needs to be larger than S_{idle}. We define the duty cycle of the PLM as the $S_{\text{min}}/S_{\text{idle}}$ ratio. The minimum duty cycle that allows the system to operate continuously depends on the implementation and may be evaluated experimentally. Moreover, $S_{\text{max}} = M \times S_{\text{min}}$ should be smaller than the fragmentation threshold in the smartphone Wi-Fi interface.

Finally, the smartphone needs to send preamble and trail Wi-Fi packets to provide the energy required for the iPoint to start-up, process the information and send the reply message back through I2S channel.

Bit Error Rate of PLM Before we start to estimate the bit error rate (BER) of the PLM, let us take a closer look at the decoding mechanism. As previously explained, messages are separated with an idle period of length $S_{\text{idle}} = S_{\text{min}}$. Let us assume that MCU records N samples during that period. The sample is considered low (OFF) if

the recorded voltage is below a threshold (V_{th}), otherwise is recorded as high (ON). PLM decoder uses a sliding window to detect the S_{idle} while counting the recorded samples. The decoder detects S_{idle} when more than $\lfloor N/2 + 1 \rfloor$ samples in the window are recorded as OFF. Once S_{idle} is detected, the number of the samples up to that point determines the preceding message. Given this decision mechanism, there are two possible scenarios that the decoder receives an incorrect message:

CASE I: If the decoder fails to detect an idle period. In this case, two messages around the undetected idle period are both lost.
CASE II: If the decoder mistakenly detects an idle period in the middle of a message. In this case, the incoming message is lost.

First, we estimate the probabilities of sample error. Because of the close proximity of the communication entities and presence of line of sight, we model the channel between source and the receiver as additive white Gaussian noise (AWGN) channel with good approximation. Therefore, the received sample, y can be expressed as $y = s + n$, where n is the noise estimated by a zero-mean normal distribution with variance of N_0. During the idle period, the input of the envelope detector is the white noise, n. If Gaussian noise is passed through an envelope detector, the probability density function (PDF) of the envelope of the noise at the output of the detector can be estimated with the following Rayleigh density function [70]:

$$P(x) = \frac{x}{N_0} e^{-x^2/2N_0} \tag{18}$$

Therefore, the probability of the error (Blue area as shown in Fig. 8) is given by:

$$P(e|s_{\text{off}}) = P(x > V_{\text{th}}) = \int_{V_{\text{th}}}^{\infty} \frac{x}{N_0} e^{-x^2/2N_0}\, dx. \tag{19}$$

When the signal is present, the PDF of the output of envelope detector can be estimated by Rician distribution [72]. Assuming that the amplitude of the signal is A, we have

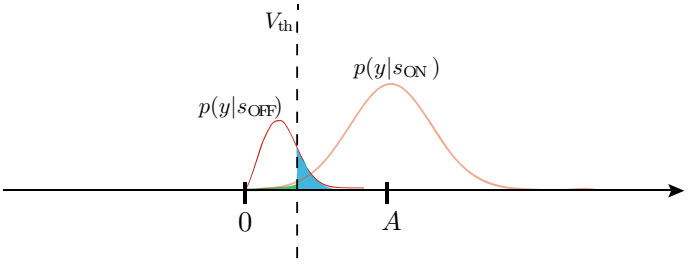


Fig. 8 Probability density function of the received sample with PLM modulation

$$P(x|A) = \frac{x}{N_0} e^{-(x^2+A^2)/2N_0} I_0\left(\frac{xA}{N_0}\right), \tag{20}$$

where I_0 is modified Bessel function of order zero. The sample error occurs when detected envelope falls below voltage threshold and has the following probability (Green area as shown in Fig. 8):

$$P(e|s_{\text{on}}) = P(x < V_{\text{th}}) = \int_0^{V_{\text{th}}} \frac{x}{N_0} e^{-(x^2+A^2)/2N_0} I_0\left(\frac{xA}{N_0}\right) dx \tag{21}$$

Also, samples are independent and identically distributed (i.i.d.). Consequently, the probability of *CASE I*, that is when more than $\lfloor N/2 + 1 \rfloor$ samples that were sent as s_{off} are detected as s_{on}, can be calculated as the following.

$$p_{\text{I}} = \sum_{i=\lfloor N/2+1 \rfloor}^{N} p(e|s_{\text{off}})^i (1 - p(e|s_{\text{off}}))^{N-i}, \tag{22}$$

where $p(e|s_{\text{off}})$ is the probability of an error given an OFF sample was transmitted. In *CASE II*, the probability of error in a message of length $m \times S_{\text{idle}}$ is:

$$p_{\text{II}}^{\text{m}} = (1 + (m-1)N) \sum_{i=\lfloor N/2+1 \rfloor}^{N} p(e|s_{\text{on}})^i (1 - p(e|s_{\text{off}}))^{N-i}. \tag{23}$$

The total probability of CASE II can be calculated by averaging over all the message sizes. For PLM with M different message, we have

$$p_{\text{II}} = (1 + \frac{(M-1)N}{2}) \sum_{i=\lfloor N/2+1 \rfloor}^{N} p(e|s_{\text{on}})^i (1 - p(e|s_{\text{off}}))^{N-i}. \tag{24}$$

We assume that the probability of sending ON and OFF messages are equal $p(ON) = p(OFF) = 0.5$. Each message codes $\log_2 M$. Therefore, the total BER can be written as:

$$BER = \frac{1}{2} \log_2 M (2p_{\text{I}} + p_{\text{II}}) \tag{25}$$

Interference The scale of the PLM performance degradation due to interference is limited because the device operates in short range. The sources of Wi-Fi interference with large output power (e.g., access points or RFID readers) are commonly located far from the device and as a result, do not pose any disruptive interference.

4.6.2 LCD Pattern Coding (LPC)

The transmitting power of the smartphone is much lower than a conventional RFID reader (few milliwatts for the smartphone versus few watts for the RFID reader). Therefore, using a similar scheme as passive RFID tags (i.e., backscattering) is not practical. Instead, we introduce LPC, a low-cost way of sending information to the smartphone taking advantage of the imaging and computing capability of the phone.

Encoding Having processed the request sent from the user, iPoint encodes the information in a series of LCD segment patterns and displays them on the panel. The smartphone captures the sequence of the patterns with the camera, recognizes the patterns, interprets the information, and finally sends the interpreted data to the user through its own UI. Because all the expensive operations are done on the smartphone side, the encoder/transmitter complexity of the iPoint may reduce significantly. This also proves to be a very energy-efficient method as displaying information on the LCD panel requires far less energy compared to conventional back scattering scheme used in passive RFID tags. To get an intuition about the energy efficiency of LCD displays, one can think of the lifespan of wrist watches with LCD display; they run for years on a tiny button cell holding a small amount of charge.

An LCD pattern is a combination of the LCD display's segments where a segment can be ON or OFF. Upon receiving a request, the MCU computes the LPC encoded output message, as a sequence of predefined patterns to be shown on the LCD panel. A two-dimensional (2D) barcode encoding such as QR codes may be used to encode the information. An LPC message that consists of n patterns on a M-segment LCD panel, encodes $n \times M$ bits of information (Fig. 9).

Decoding At the other end of the channel, the LPC message is captured by the smartphone camera by either recording a video or taking a series of pictures at a satisfying rate. The smartphone decodes the captured message by running a pattern recognition algorithm on each frame, and sends the interpreted data to the user via UI or uses the data in the next sessions of communication. Several 2D barcode decoding algorithm and software for smartphones have been published [24]. An example of such setting is shown in Fig. 10.

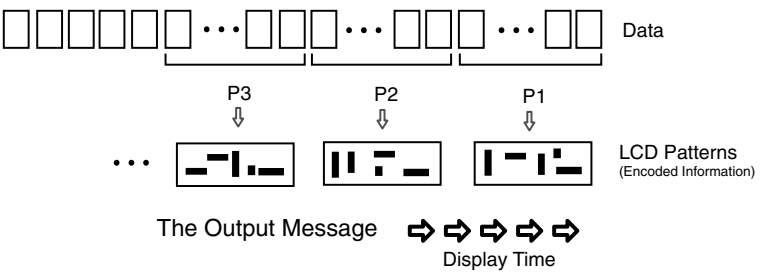


Fig. 9 LPC Encoding for a M-segment LCD panel

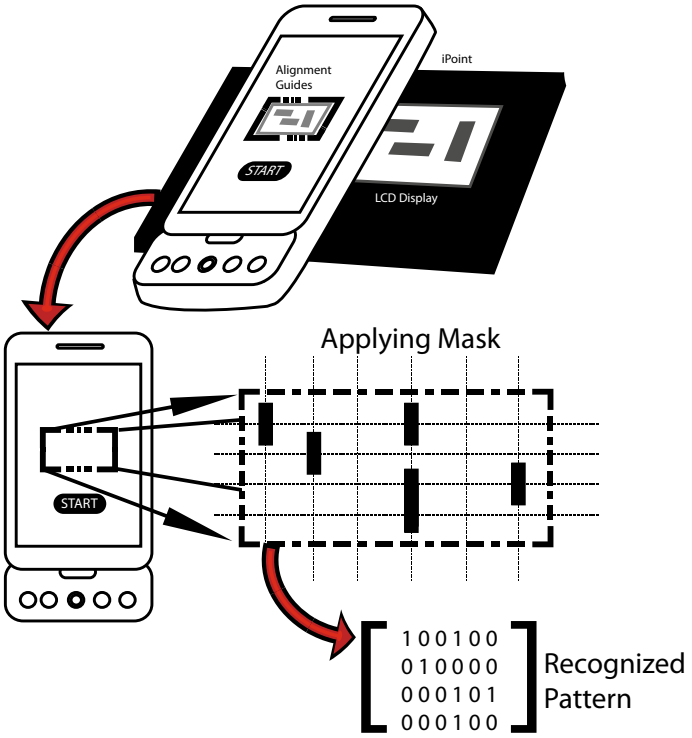


Fig. 10 LPC decoding. To make decoding faster, the user is asked to align the image of the panel within a virtual box, then the frame is sampled only on the intersections of mask grid lines (Narrow dashed lines)

LPC Rate Analysis: Using passive-matrix panels, the LCD pattern update rate can go up to 200 Hz. However, our experiments indicate that the deciding factor on an error-free decoding is the sampling rate of the camera. Let R_c denote the maximum sampling rate of the smartphone camera. Applying Nyquist theorem, we have, $R_{max} < R_c/2$ fps. Therefore, for a M-segment LCD display, we have the upper bound transmission rate of $R_c M/2$ bps. For most of today's commercial smartphone cameras, $R_c \approx 30$ fps.

4.7 Optimization Techniques

In this section, we take a closer look into the techniques to optimize the performance of the iPoint. We review the design choices for each component of the system, and present advantages and disadvantages for each design. Also, we study the trade-offs that emerge in the course of the hardware and software design.

4.7.1 Antenna

The antenna is a key component of the iPoint design and its performance has a major impact on the overall efficiency of the system. Note that in any communication system, the design of both receiver and transmitter antenna should be optimized. However, in order to maintain universality in our design, we do not have any control over the transmitting antenna on the smartphone. Therefore, we assume that the smartphone is equipped with an omnidirectional antenna optimized for the Wi-Fi frequency and focus on improving the receiving antenna. In the design of iPoint, we aim to maximize the gain of the antenna as well as its radiation performance over the operating frequency band, which is 2.4 GHz for iPoint. The cost and size of the antenna plays an important role in choosing the proper antenna design. There are three main types of antennas that can be used in our design: whip antennas, chip antennas, and PCB planar antennas. Whip antennas consist of a single straight piece of conductor, normally in the form of a wire, mounted over the ground plane. These antennas have the simplest design and can achieve high gains (up to ≈ 6 dBi). However, they should extend perpendicular to the ground plain and the board to achieve the best performance, hence, are not easily fit in compact board designs. Chip antennas are smaller and can be easily mounted on electronic boards, but have a radiation performance ranging from mediocre to poor. PCB antennas provide the most flexibility in terms of antenna geometry. They are planar and have small size in high frequencies therefore can fit in compact designs. The RF performance of the PCB antennas are typically worse than whip antennas, but a good performance can be achieved with a careful design using microwave simulation tools. In the following, we present three PCB antennas that we designed for iPoint. We list the characteristics, advantages, and disadvantages of each design.

PCB Dipole Antenna Dipole antennas have been widely used since the early days of radio. Simplicity and effectiveness for a wide range of communication needs are the reasons for this. A dipole, which it gets its name from its two halves, is a balanced antenna, meaning that the poles are symmetrical: they have equal lengths and are extended in opposite directions from the feed point. To be resonant, a dipole must be electrically a half wavelength long at the operating frequency. Figure 11 shows the geometry of our dipole antenna which is designed to operate at 2.4 GHz. The return loss S_{11} of this antenna is presented in Fig. 11.

Yagi–Uda Antennas Higher gain antennas are usually obtained by forming arrays of basic antennas. The Yagi–Uda antenna is the most successful general-purpose directional antenna design at frequencies up to 2.5 GHz. It is inexpensive and simple to construct, and will provide gains of up to about 17 dBi. Yagi–Uda antennas can be built to support high input powers, and they are commonly used for directional broadcast transmission. The geometry of the designed Yagi–Uda antenna, which is operating at 2.4 GHz, is shown in Fig. 12. The return loss of this antenna is also illustrated in Fig. 12. The gain and directionality of Yagi–Uda antennas are particularly desirable. However, their large size in our operating frequency becomes problematic and renders their use impractical.

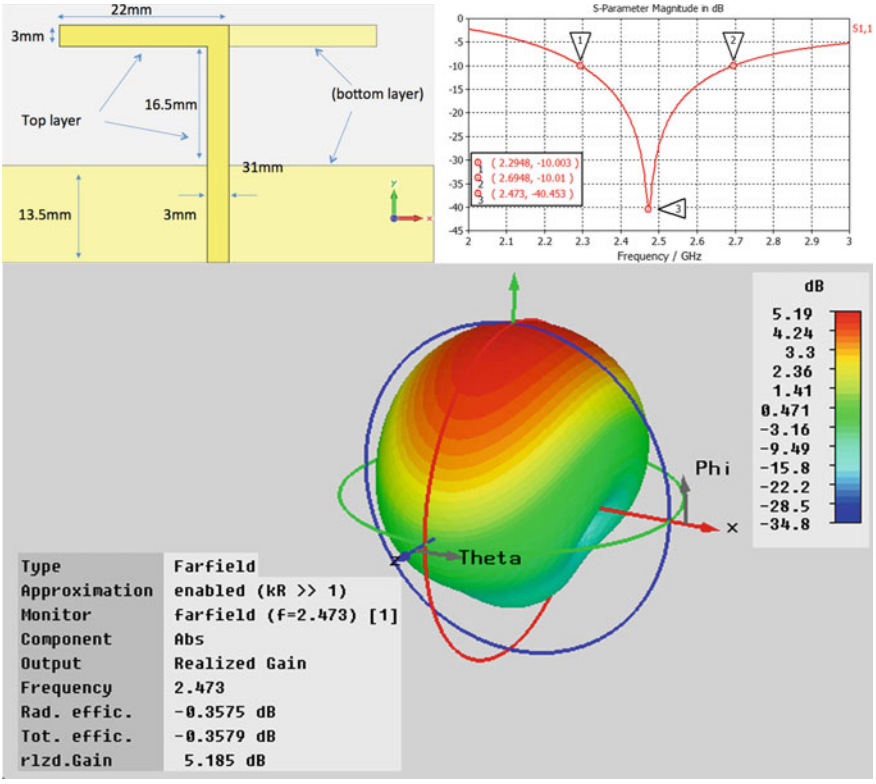


Fig. 11 The design of our dipole antenna optimized for 2.4 GHz [37]

Planar Inverted-F Antenna The planar inverted-F antennas (PIFA) are commonly used in mobile communication devices due to their small size (quarter-wavelength). These antennas typically consist of a rectangular planar element located above a ground plane, a short-circuiting plate, and a feeding mechanism for the planar element. The Inverted-F antenna is a variant of the monopole where the top section has been folded down so as to be parallel with the ground plane. This is done to reduce the height of the antenna, while maintaining a resonant trace length. This parallel section introduces capacitance to the input impedance of the antenna, which is compensated by implementing a short-circuit stub. The stub's end is connected to the ground plane through a via connection. Figure 13 shows the geometry of our PIFA antenna designed to operate at 2.4 GHz [37]. The return loss of this antenna is also shown in Fig. 13. PIFA's size makes it a very good choice for compact board design, however, its performance is subpar compared to dipole and Yagi–Uda antennas.

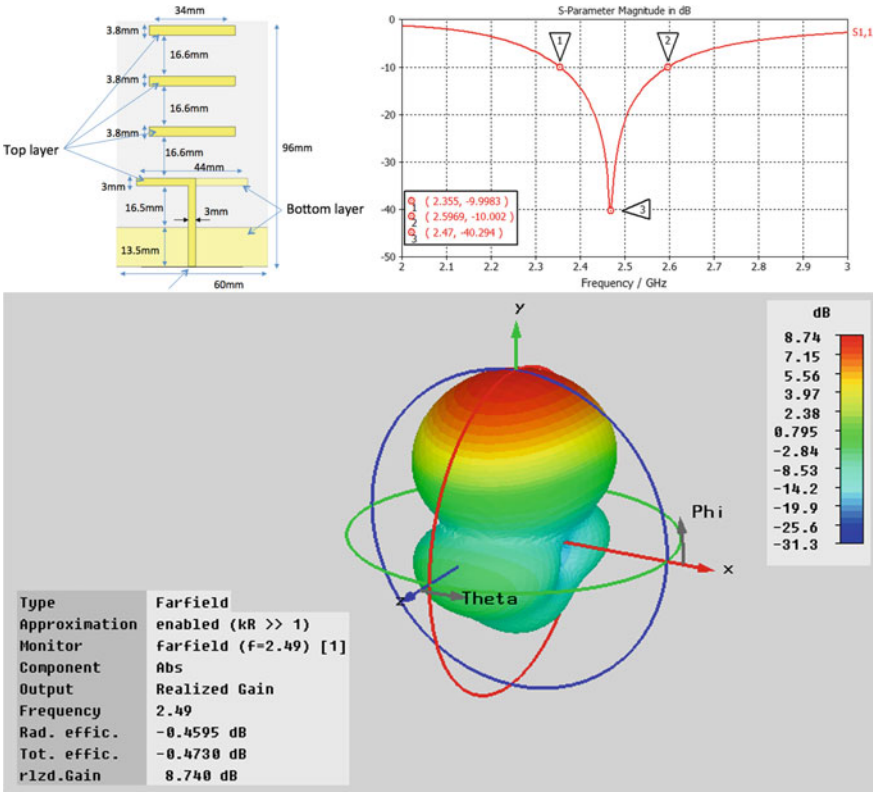


Fig. 12 The design of the Yagi–Uda antenna array optimized for 2.4 GHz

4.7.2 Rectifier

The rectifier designed for iPoint is based on multi-stage Greinacher circuit described in Sect. 3.2.4. The overall efficiency of the rectifier is determined by combination of the design of the rectifier and the electrical characteristics of its components. In this section, we review the key elements in designing an optimized rectifier for iPoint.

Diodes Since the rectifier is used to convert high-frequency Wi-Fi signals, the diode used in the design should have fast switching times. Among available types of diodes, Schottky diodes provide the fastest switching time, therefore are most suitable for RF energy harvesting. Another benefit of using Schottky diodes is their low forward voltage (150–350 mV), which is considerably lower than normal p–n junction diodes (0.6 V). The lower forward voltage allows a greater portion of the signal to be rectified and directly improves the efficiency of the rectifier. Other important parameters that need to be considered in picking the proper diode for the design is the saturation current, junction capacitance, and series resistance. In particular, we are looking for the following characteristics:

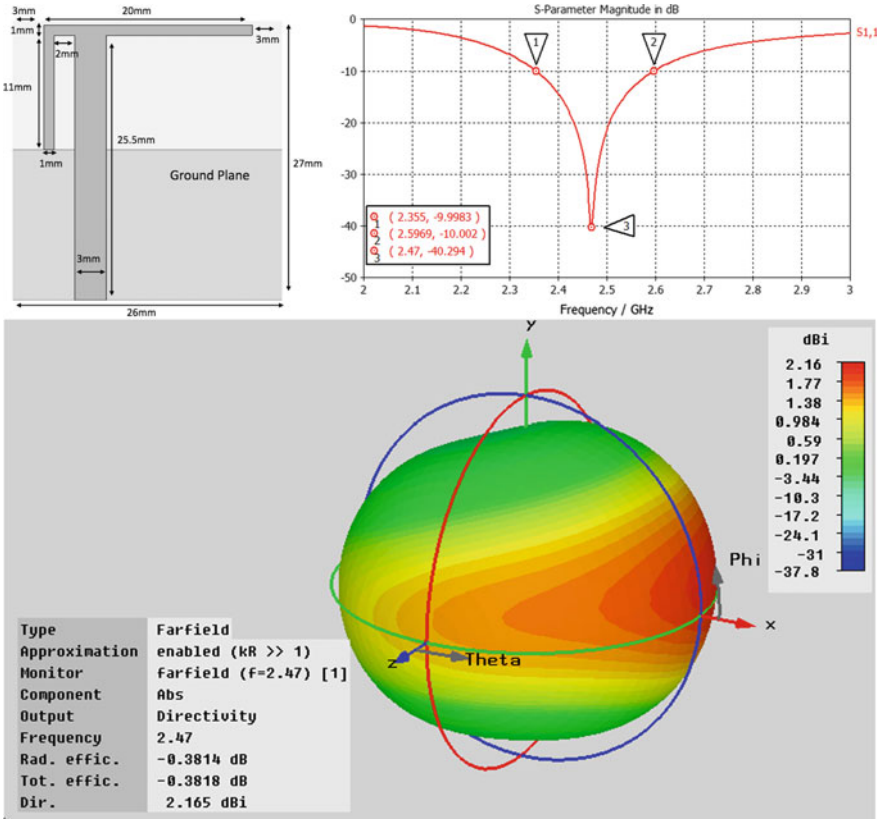


Fig. 13 The design of the Planar Inverted-F Antenna optimized for 2.4 GHz

- High saturation current is required for driving heavier loads.
- Lower junction capacitance leads to lower overall rectifier reactance and reduces the frequency dependence of the rectifier, and consequently eases the impedance matching.
- Lower series resistance is desirable in order to reduce the power loss within the rectifier.

In our prototype, we use HSMS-282x Schottky diodes from Avago Technologies since they provide the best combination of the electrical characteristics within the operating frequency and power band.

Load Impedance As described before, the load impedance greatly affects the performance of the rectifier. The impedance of the MCU computing core of the iPoint varies depending on the state of the computation. The low-power mode (LPM) of the MCU shows the highest impedance, whereas the lowest impedance was measured when MCU was driving the LCD. The highest power consumption coincides with

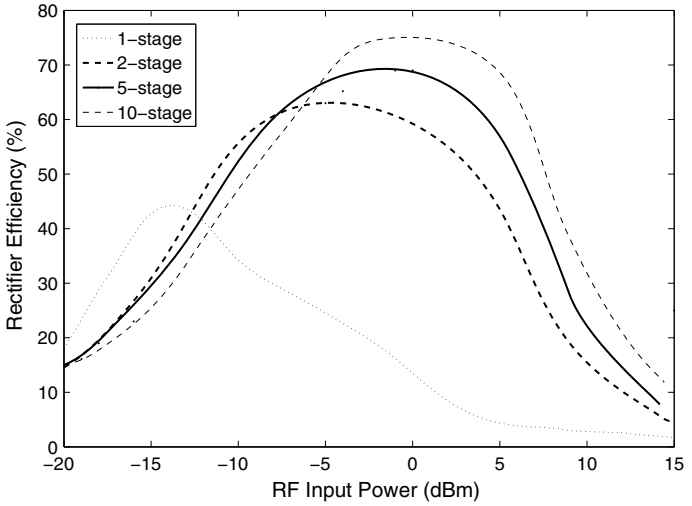


Fig. 14 Rectifier efficiency versus Input power for different number of the stages

the lowest impedance of the MCU, measured as 140 kΩ; In order to ensure that the rectifier provides necessary power at all times, we modify the rectifier to have the best efficiency while driving the maximum load.

Rectifier Topology The number of the stages and RF input of the rectifier have significant effect on the output voltage level and efficiency of the rectifier.
Number of Stages In an ideal multi-stage rectifier, the output voltage is proportional to number of the stages. Nevertheless, that does not hold in practice. The parasitic effects of the intermediate stages plus the power loss due to resistance in each stage limits the practical number of stages. Obtaining a closed-form analytical solution in a multi-stage rectifier is not practical, but the optimization can be done using realistic simulation. We used Agilent ADS tool to simulate rectifiers with different number of stages. The components of each circuit are identical and simulated using the vendor-provided parameters. Figure 14 shows the efficiency of the rectifiers versus input RF power. It shows that the maximum efficiency of the rectifier improves as number of stages grows. The simulation result also shows that rectifiers with higher number of stages show better efficiency at higher input powers.

4.7.3 Matching Strategy

The goal of the matching network is to adjust the input impedance of the system seen from the antenna-rectifier interface, Z_{in}, in order to have

$$Z_{\text{in}} = Z_{\text{ant}}^{*}. \tag{26}$$

Here, lies a peculiar problem: the diodes are nonlinear electrical components, hence the rectifier circuit shows a collective nonlinear behavior. This implies that Z_{in} depends on the input power from antenna, P_{R}. This is not an ideal situation because P_{R} may vary unpredictably due to numerous dependencies discussed in Sect. 3.2.2. The dynamic input impedance calls for a dynamic and adaptive matching mechanism that cannot be achieved using passive components.

A viable solution to minimize the effect of impedance mismatch is to measure the variation of Z_{in} for a reasonable range of input power and design the matching circuit accordingly. In the case of iPoint, we assume that the smartphone is always transmitting Wi-Fi signals at the maximum power allowed by regulations. Therefore, the optimal matching occurs when the communication occurs in a specific distance, which can be predicted.

The problem of dynamic impedance remains even if the input power is constant. This is because the input impedance of the system is also affected by the current that computing unit draws from the rectenna, i_{out}. For example, the MCU requires significantly higher current in the active mode (i.e., performing computation) compared to standby (i.e., low-power mode), which results in lower input impedance. If matching is optimized for the standby mode, any increase in i_{out} results in quick drop of output voltage of the rectifier leading to system shutdown. A partial solution is to match the rectenna for an desired input impedance, which in most cases is minimum value of Z_{in} when $i_{\text{out}} = max$. Given a fixed input power, when i_{out} falls below maximum value, a power mismatch occurs but in fact the output voltage of the rectifier increases. This prevents the system from shutting down as long as the input power is sufficient.

In our design, matching is done using an LC network plugged between the antenna and the rectifier. The final trade-off in matching is the operating bandwidth of the system. While high-quality sharp matching maximizes the power transfer efficiency, it also makes the system more frequency selective. This might not be ideal where the operating bandwidth of the system is not small. The fundamental limit of impedance matching over a bandwidth can be estimated by Bode-Fano theorem. Let us assume the rectifier circuit can be represented by a parallel RC network. According to Bode and Fano, if a lossless matching network is employed, we have the following limit on the reflection coefficient for different frequencies:

$$\int_0^\infty \ln\left(\frac{1}{|\Gamma|}\right) \mathrm{d}\omega \leq \frac{\pi}{RC}. \tag{27}$$

A perfect matching over a specific bandwidth $\Delta\omega = \omega_2 - \omega_1$ is theoretically achieved if $\Gamma = 1$ for frequencies outside the bandwidth and $\Gamma = \Gamma_{\text{min}}$ for $\omega_1 < \omega < \omega_2$. Equation 27 gives the following lower bound for operating reflection coefficient:

$$\Gamma_{\text{min}} > e^{-\pi/RC\Delta\omega}. \tag{28}$$

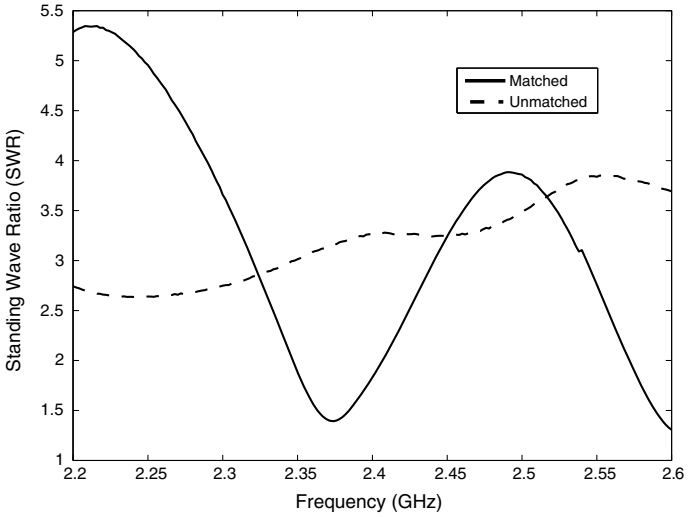


Fig. 15 The impedance matching in action. The standing-wave ratio (SWR) of the rectenna is shown before and after careful matching at active mode load

Dynamic impedance of the system makes the matching mechanism a very complex problem. As previously explained, the efficiency of the rectifier is also dynamic and shares some of the same dependencies. In practice, it is necessary to determine the value of matching components by optimizing the output power of the rectenna, which accounts for both matching and rectifying efficiency. Figure 15 shows the performance of the designed matching network for iPoint's rectenna.

4.7.4 Low-Power Computation

Power-aware Software The embedded software of the iPoint takes advantage of low-power modes (LPM) provided by MCU to minimize the power consumption. Any MCU components (timer, ADC, etc.,) can be turned off when its functionality is not required.

Underclocking To further reduce the power consumption of the computing core, we aggressively underclock the MCU. The iPoint's computing core tasks, such as PLM decoding and LPC encoding, do not demand a very fast clock, hence the clock frequency may be reduced to a few kilohertzs. Figure 16 illustrates the results of our measurements of the MCU's power consumption running the same instructions at different clock frequencies.

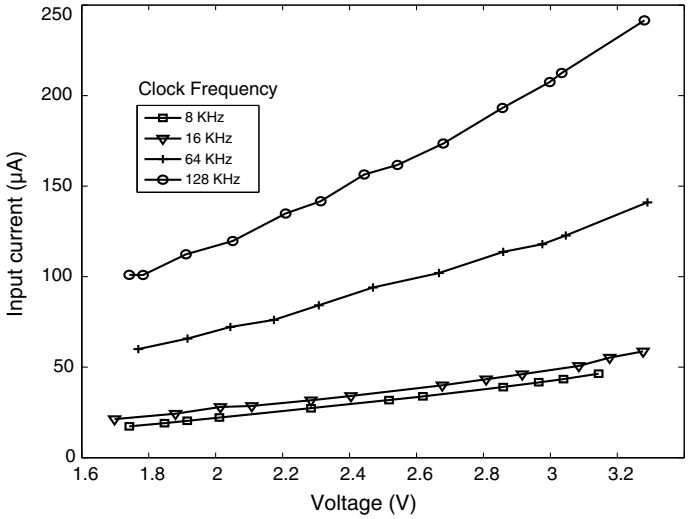


Fig. 16 Power consumption of MCU for different clock frequencies

4.7.5 Prototype

We prototyped different versions of the iPoint based on the design described in Sect. 4.5. Two versions (Ver. 2.1 and 3.1) are shown in Fig. 17. In this section, we explain the implementation of the components in detail.

The smartphone used in the experiments is the *HTC Dream* also known as *T-Mobile G1* running Android mobile device platform Ver. 1.6. This 3G phone is equipped with a 528 MHz Qualcomm ARM11 Processor, 192 MB of DDR SDRAM, 320×480 pixel LCD Display with 180 ppi, 3.2-megapixel camera with auto-focus capability, and a Wi-Fi (802.11 b/g) wireless interface [35]. We developed a software application in Android platform that sends multiple PLM-modulated requests, and performs the LPC decoding. The Wi-Fi interface was configured to send broadcast packets at a fixed rate of 1 Mbps, the lowest rate supported by Wi-Fi communication. Ideally, the application should create an ad hoc network, but the Android's support for the ad hoc mode is currently limited. As an alternative solution for prototyping, the smartphone connects to an auxiliary Wi-Fi network created by an external access point. We use UDP/IP, as opposed to TCP/IP, to avoid unnecessary retransmissions caused by TCP flow control mechanism. For the iPoint rectenna, we implemented a ten-stage modified Greinacher circuit, a full-wave rectifier with parallel RF inputs connected to a 2.4 GHz whip antenna. We used high-performance low-leakage RF capacitors, and Schottky diodes (HSMS-282 series from Avago technologies) with forward voltage threshold of 150–200 mV, the lowest available. The value of intermediate capacitors were chosen experimentally to maximize the output DC voltage. The rectenna then was matched on Wi-Fi channel 1 (2.412 GHz) using an LC matching network. The first stage of the rectifier circuit was used as an envelope detector

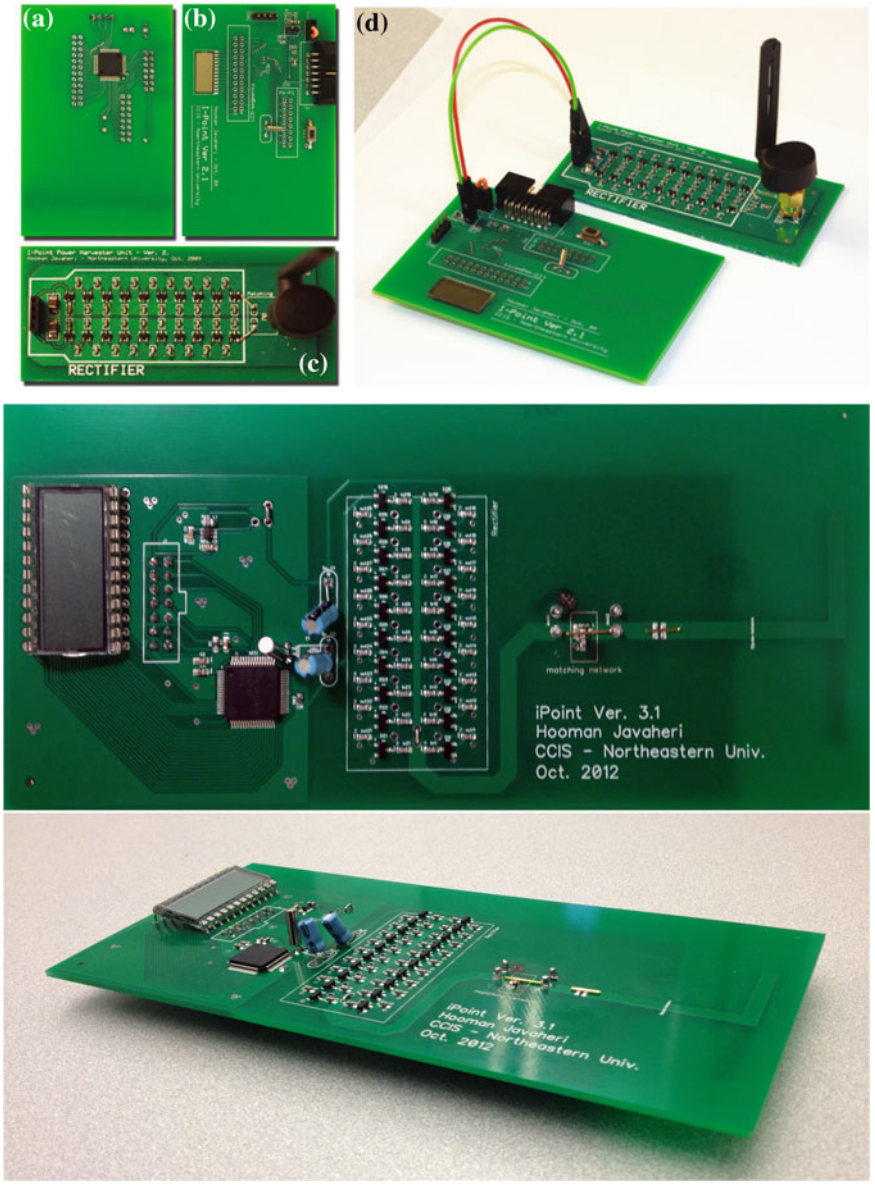


Fig. 17 The Prototype of iPoint. *Top*: The iPoint prototype boards: **a** Computing core upper layer (LCD display is shown). **b** Computing core lower layer (MSP430 is shown). **c** RF energy-harvester front end, ten-stage Greinacher voltage multiplier. **d** The realization of iPoint version 2.1. *Middle*: Prototype Version 3.1 with unified board design and PCB antenna. *Bottom*: Low-profile design of the iPoint Ver. 3.1

circuit for PLM decoding. For the communication core, we embed a TI MSP430F417, an ultra-low-power microcontroller from Texas Instruments. This 16-bit flash MCU provides desired computing capabilities at low-power consumption. It features 32 kB + 256 B of flash memory, 1 kB of RAM, Low supply voltage of 1.8 V, integrated LCD driver for 96 segments, on-chip comparator that can be used for finalizing the PLM signal demodulation, and very low active power consumption of 200 μA at 1 MHz, which makes it a reasonable choice for the iPoint prototype. The LCD panel selected for this generation of the prototype was a 26 segment watch LCD display.

4.7.6 Performance Evaluation

We carried out several experimental measurements in order to accurately characterize the device and prove the functionality of the design components. This section presents the detailed description of the testbed and experimental results.

Range Based on our experiment, iPoint requires around −10 dBm of power to operate normally. The transmit power of Wi-Fi interface of smartphones varies from a model to another but can be roughly estimated between 10–20 dBm. If the antennas at both receiver and transmitter side provide gain of $G_T = G_R \approx 3$ dB, and we have polarization mismatch of $L \approx -3$ dB, the link budget analysis gives an operating range up to two times of the signal wavelength, which is 12.5 cm for Wi-Fi signal. Our experimental evaluation is in agreement with the preceding estimation; our prototype was fully operational in ranges below 25 cm.

Rectifier Efficiency To characterize the efficiency of the rectifier circuit, a MXG Vector Signal Generator was used to feed the rectifier via a 0.5 ft coaxial cable, and the output voltage level of the rectenna was measured. The rectifier was fed with a Wi-Fi signal in a wide range of input power, from −20 dBm to 15 dBm. The output voltage and efficiency were measured without a load and with a load of 140 kΩ, which is close to the MCU impedance in active mode. Rectifier shows efficiency levels up to 72%. The results are shown in Figs. 18 and 19.

Duty Cycle of PLM Minimum duty cycle as one of the important characteristics of the iPoint system was discussed in Sect. 4.5. We measured the output DC voltage of the rectenna for different duty cycles. The results are summarized in Fig. 20.

LCD Contrast Test The accuracy of LPC decoding relies on the contrast of the pattern shown on the LCD panel, light conditions and the distance of the camera from the panel. If available on the smartphone, an LED flash may be used to compensate low-light conditions.

The power consumption of two LCD panels with different sizes were measured: Panel 1 (3 cm^2, 24 segment) and Panel 2 (1.8 cm^2, 26 Segments). A test image was taken from the LCD panels at the same distance and under the same light environment while the same pattern were displayed on both panels. The contrast of the panels were compared digitally in Adobe Photoshop. To create a given desired contrast, we measured the required voltage and input current for each panel. The larger panel,

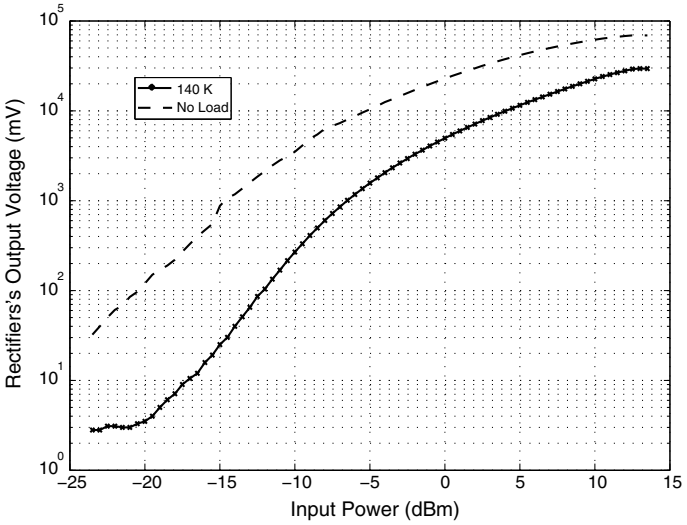


Fig. 18 Performance of the energy-harvester unit

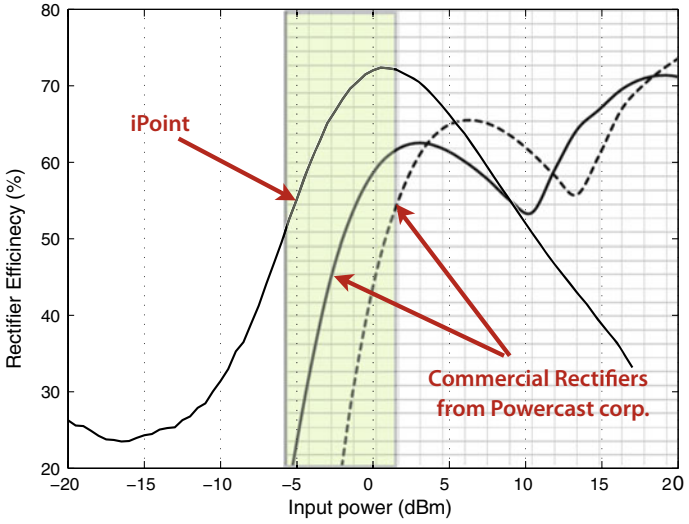


Fig. 19 Energy-harvester efficiency as a function of input power. The shaded region indicates the operating range of the iPoint. Based on our measurements the optimized rectifier in our prototype outperforms state-of-the-art commercial rectifier chips within desired input power range

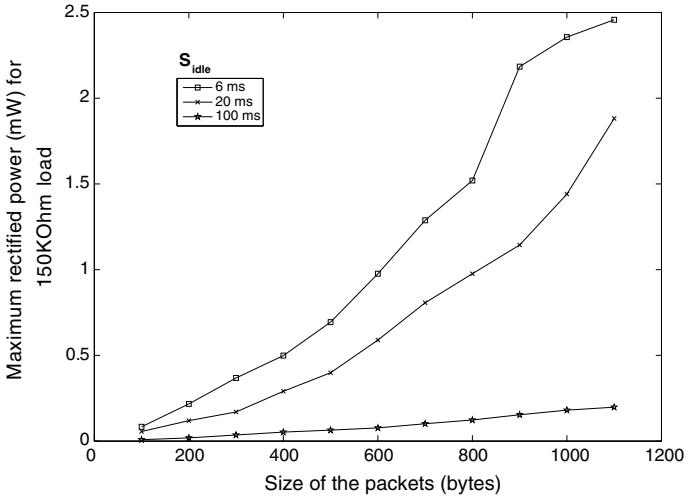


Fig. 20 The rectified output voltage as a function of the packet length for different idle times S_{idle} (therefore duty cycles)

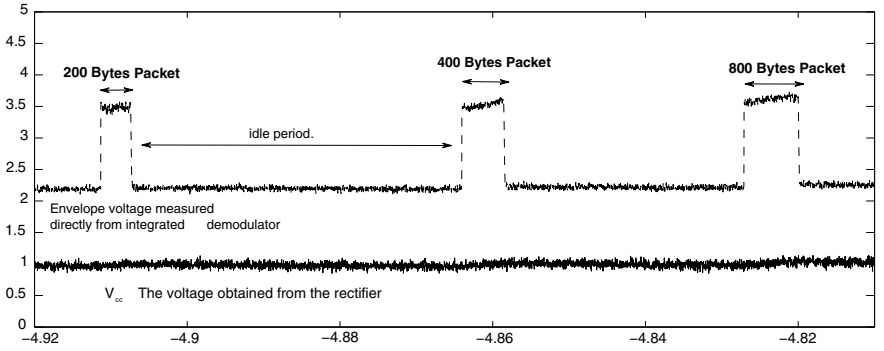


Fig. 21 The voltage of energy harvester V_{cc} and integrated demodulator (envelope signal) in Volts. Note that, the size of the packet is easily detectable. The Wi-Fi communication rate was fixed to 1 Mbps. harvester's output level is fairly smaller than the peaks of envelope signal. That reason is the long idle times between packet transmissions. The data is captured by a Infinium MSO8104 oscilloscope from Agilent Technologies. The plot is regenerated in MATLAB

requires 2.9 V drawing 49 μA (142.1 μW), whereas Panel 2 requires 1.9 V drawing 21 μA (39.9 μW).

PLM Decoding Performance In order to test the functionality of the integrated demodulator of the front end, packets with different sizes were sent over the Wi-Fi channel by the smartphone while the output of the energy harvester and integrated demodulator being measured. The rate of communication was fixed to 1 Mbps. The result is shown in Fig. 21.

4.8 Summary

In this section, we introduced iPoint, a device that can interact with commodity smartphones, equipped with a Wi-Fi network interface and camera, therefore enabling ad hoc and universal communication. The energy and information are exchanged in an instance of a previously described CEICh channel. At the core of iPoint lies an ultra-low-power microcontroller. iPoint draws the entire of its energy from the smartphone transmissions, through an RF energy-harvester, making the use of batteries unnecessary and guaranteeing its longevity. Two new communication paradigms are introduced:

- Packet width modulation allows the smartphone to encode information in the width of the Wi-Fi packet and making demodulation extremely energy-efficient
- LCD information encoding and camera decoding.

We discussed several design possibilities and built a prototype of the iPoint. We reported on the performance of our system for various transmission powers, operation frequencies of the microcontroller, packet sizes and duty cycles.

5 Conclusion and Future Research Directions

Energy efficiency of wireless communications remains a key obstacle that greatly influences the overall performance of wireless networks. Critical dependency of wireless devices on their limited source of energy requires careful adjustments in computations and communication to conserve as much energy as possible. In some particular applications of wireless communication, the energy conservation issue becomes so severe that having a local source of energy (battery) is not practical or feasible. One example of such scenario can be observed in wireless sensor networks in which the sensor node consumes a vast majority of its energy on keeping its wireless radio on waiting for a typically rare incoming message. In this work, we explore the wireless transfer of energy alongside information as a solution to energy conserving problem in the aforementioned scenarios.

We have presented the notion of consolidated energy and information channel for wireless sensor networks. We argue that energy conservation issues at the receiver side can be eliminated by transferring energy via wireless signals. We have reviewed the steps required to enable such functionality in wireless sensor networks, studied potential wireless energy transfer methods, and presented the necessary modifications to wireless communication after the integration of energy transfer. We have introduced iPoint, a passively powered wireless device that is capable of communicating with a commodity smartphone without any need for a battery or any specific hardware modification on the phone. The complete design and optimization of the software and hardware of the device were presented. We presented two new communication schemes: packet length modulation (PLM) and LCD pattern coding (LPC).

We provided the theoretical performance analysis, and also performed a rigorous experimental evaluation of the system. iPoint is an example of a wireless device that consumes no energy unless it is necessary. At first glance, this seems similar to the functionality of RFID passive tags. However, iPoint's ability to communicate with a device as common as a smartphone without any hardware modification (just by installing an application) makes the information much more accessible. The design of the passively powered devices is a vastly complex problem. While we tried to optimize the different components of the design as much as possible, there is still room for improvement. In the future research, it would be interesting to consider more sophisticated antenna designs to improve the energy harvesting performance. Another interesting problem is to study the effects of installing security schemes on top of the presented communication protocols on the overall energy efficiency of the system.

References

1. Adair, R.K.: Constraints on biological effects of weak extremely-low-frequency electromagnetic fields. Phys. Rev. A **43**(2), 1039–1048 (1991)
2. Adair, R.K.: Vibrational resonances in biological systems at microwave frequencies. Biophys J **82**(3), 1147–52 (2002)
3. Alphandéry, E., Faure, S., Raison, L., Duguet, E., Howse, P.A., Bazylinski, D.A.: Heat production by bacterial magnetosomes exposed to an oscillating magnetic field. J. Phys. Chem. C **115**(1), 18–22 (2011)
4. WiTricity Corp. http://www.witricity.com/pages/technology.html
5. Del Barco, E., Asenjo, J., Zhang, X., Pieczynski, R., Julia, A., Tejada, J., Ziolo, R.F., Fiorani, D., Testa, A.M.: Free rotation of magnetic nanoparticles in a solid matrix. Chem. Mater. **13**(5), 1487–1490 (2001)
6. Bennett, C.: Logical reversibility of computation. IBM J. Res. Dev. **17**(6), 525–536 (1973)
7. Berridge, M.:The AM and FM of calcium signalling. Nature (1997)
8. Blakemore, R.: Magnetotactic bacteria. Science **190**(4212), 377–379 (1975)
9. Buettner, M., Prasad, R., Sample, A., Yeager, D., Greenstein, B., Smith, J.R., Wetherall, D.: Rfid Sensor Networks with the Intel Wisp. pp. 393–394 (2008)
10. Blatt, J.M., Weisskopf, V.F.: Theoretical Nuclear Physics, p. 864 (1954)
11. Chen, C.: Remote control of living cells. Nature Nanotechnology (2008)
12. Cifra, M.: Electrodynamic eigenmodes in cellular morphology. BioSystems **109**(3), 356–66 (2012)
13. Cohen, R., Kapchits, B.: An optimal wake-up scheduling algorithm for minimizing energy consumption while limiting maximum delay in a mesh sensor network. IEEE/ACM Trans. Netw. **17**(2), 570–581 (2009)
14. Chaubey, A., Malhotra, B.D.: Mediated biosensors. Biosens. Bioelectron. **17**(6–7), 441–456 (2002)
15. Choi, J.H., Nguyen, F.T., Barone, P.W., Heller, D.A., Moll, A.E., Patel, D., Boppart, S.A., Strano, M.S.: Multimodal biomedical imaging with asymmetric single-walled carbon nanotube/iron oxide nanoparticle complexes. Nano Lett. **7**(4), 861–867 (2007). Apr
16. Cockcroft, J.D., Walton, E.T.S.: Experiments with high velocity positive ions. (i) further developments in the method of obtaining high velocity positive ions. Proc. R Soc. Lond. Ser. A **136**(830), 619–630 (1932)

17. Correia, L.M., Zeller, D., Blume, O., Ferling, D., Jading, Y., Góanddor, I., Auer, G., Van Der Perre, L.: Challenges and enabling technologies for energy aware mobile radio networks. IEEE Commun. Mag. **48**(11):66–72 (2010)
18. Dykman, M.I., Khasin, M., Portman, J., Shaw, S.W.: Spectrum of an oscillator with jumping frequency and the interference of partial susceptibilities. Phys. Rev. Lett. **105**(23), 230601 (2010)
19. Dobson, J.: Remote control of cellular behaviour with magnetic nanoparticles. Nat. Nanotechnol. (2008)
20. Degen, C.L., Poggio, M., Mamin, H.J., Rettner, C.T., Rugar, D.: Nanoscale magnetic resonance imaging. Proc. Natl. Acad Sci. USA **106**(5), 1313–1317 (2009)
21. Drubach, D.: The Brain Explained, p. 168, Jan 2000
22. Ermolov, V., Heino, M., Karkkainen, A., Lehtiniemi, R., Nefedov, N., Pasanen, P., Radivojevic, Z., Rouvala, M., Ryhanen, T., Seppala, E., Uusitalo, M.: Significance of nanotechnology for future wireless devices and communications. In: 2007 IEEE 18th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC 2007, pp. 1–5 (2007)
23. EPCglobal Standards and Technology. http://www.epcglobalinc.org/standards (2008)
24. Falas, T., Kashani, H.: Two-dimensional Bar-code Decoding With Camera-equipped Mobile Phones, pp. 597–600, Mar 2007
25. Glogauer, M., Ferrier, J., McCulloch, C.A.: Magnetic fields applied to collagen-coated ferric oxide beads induce stretch-activated ca2+ flux in fibroblasts. Am. J. Physiol. **269**(5 Pt 1), C1093–104 (1995)
26. Havelka, D., Cifra, M., Kučera, O., Pokorný, J., Vrba, J.: High-frequency electric field and radiation characteristics of cellular microtubule network. J. Theor. Biol. **286**(1), 31–40 (2011)
27. Hu, X., Cebe, P., Weiss, A.S., Omenetto, F., Kaplan, D.L.: Protein-based composite materials. Mat. Today **15**(5), 208–215 (2012)
28. Hergt, R., Dutz, S., Müller, R., Zeisberger, M.: Magnetic particle hyperthermia: nanoparticle magnetism and materials development for cancer therapy. J. Phys. Condens. Matter **18**, S2919 (2006)
29. Huang, H., Delikanli, S., Zeng, H., Ferkey, D.M., Pralle, A.: Remote control of ion channels and neurons through magnetic-field heating of nanoparticles. Nat. Nanotechno. **5**(8), 602–606 (2010)
30. Howard, J., Hudspeth, A.J.: Compliance of the hair bundle associated with gating of mechanoelectrical transduction channels in the bullfrog's saccular hair cell. Neuron **1**(3), 189–99 (1988)
31. Hughes, S., El Haj, A.J., Dobson, J.: Magnetic micro- and nanoparticle mediated activation of mechanosensitive ion channels. Med. Eng. Phys. **27**(9), 754–62 (2005)
32. Hamam, R.E., Karalis, A., Joannopoulos, J.D., Soljačić, M.: Coupled-mode theory for general free-space resonant scattering of waves. Phys. Rev. A **75**(5), 53801 (2007)
33. Hughes, S., McBain, S., Dobson, J., El Haj, A.J.: Selective activation of mechanosensitive ion channels using magnetic particles. J. R. Soc. Interface **5**(25), 855–63 (2008)
34. Iyer, V., Talla, V., Kellogg, B., Gollakota, S., Smith, J.: Inter-technology backscatter: towards internet connectivity for implanted devices. In: Proceedings of the 2016 ACM SIGCOMM Conference (SIGCOMM '16), pp. 356–369. ACM, New York, NY, USA
35. HTC dream (T-mobile g1). http://www.htc.com/www/product/dream/overview.html
36. Jackson, J.D.: Classical Electrodynamics (1967)
37. Javaheri, H.: Wireless Transfer of Energy Alongside Information: From Wireless Sensor Networks to Bio-Enabled Wireless Networks. Ph.D. Dissertation. Northeastern Univ., Boston, MA, USA, Dec 2012
38. Javaheri, H., Barbiellini, B., Noubir, G.: Efficient magnetic torque transduction in biological environments using tunable nanomechanical resonators. In: 2011 Proceedings of the IEEE EMBC (2011)
39. Javaheri, H., Barbiellini, B., Noubir, G.: Efficient magnetic torque transduction in biological environments using tunable nanomechanical resonators. Conf. Proc. IEEE Eng. Med. Biol. Soc. **2011**, 1863–6 (2011)

40. Javaheri, H., Barbiellini, B., Noubir, G.: On the energy transfer performance of mechanical nanoresonators coupled with electromagnetic fields. cond-mat.mes-hall, Aug 2011
41. Javaheri, H., Barbiellini, B., Noubir, G.: On the energy transfer performance of mechanical nanoresonators coupled with electromagnetic fields. Nanoscale Res. Lett. **7**(1), 572 (2012)
42. Javaheri, H., Noubir, G.: ipoint: A platform-independent passive information kiosk for cell phones. In: 2010 7th Annual IEEE Communications Society Conference onSensor Mesh and Ad Hoc Communications and Networks (SECON), pp. 1–9 (2010)
43. Javaheri, H., Noubir, G., Noubir, S.: Rf control of biological systems: Applications to wireless sensor networks. Nano-Net (2009)
44. Janssen, X.J.A., Schellekens, A.J., van Ommering, K., van IJzendoorn, L.J., Prins, M.W.J.: Controlled torque on superparamagnetic beads for functional biosensors. Biosens. Bioelectron. **24**(7), 1937–1941 (2009)
45. Jensen, K., Weldon, J., Garcia, H., Zettl, A.: Nanotube radio. Nano letters **7**(11), 3508–3511 (2007)
46. Kirschvink, J.L.: Comment on constraints on biological effects of weak extremely-low-frequency electromagnetic fields. Phys. Rev. A (1992)
47. Karalis, A., Joannopoulos, J.D., Soljacic, M.: Efficient wireless non-radiative mid-range energy transfer. Ann. Phys. **323**(1), 34–48 (2008)
48. Kobayashi, A., Kirschvink, J.L.: Magnetoreception and electromagnetic field effects: sensory perception of the geomagnetic field in animals and humans. ACS Adv. Chem. Ser. **250**, 367–394 (1995)
49. Kirschvink, J.L., Kobayashi-Kirschvink, A., Diaz-Ricci, J.C., Kirschvink, S.J.: Magnetite in human tissues: a mechanism for the biological effects of weak elf magnetic fields. Bioelectromagn. Suppl **1**, 101–13 (1992)
50. Kurs, A., Karalis, A., Moffatt, R., Joannopoulos, J.D., Fisher, P., Soljacic, M.: Wireless power transfer via strongly coupled magnetic resonances. Science **317**(5834), 83 (2007)
51. Kim, D.H., Rozhkova, E.A., Ulasov, I.V., Bader, S.D., Rajh, T., Lesniak, M.S., Novosad, V.: Biofunctionalized magnetic-vortex microdiscs for targeted cancer-cell destruction. Nat. Mater. (2009)
52. Kirschvink, J.L., Winklhofer, M., Walker, M.M.: Biophysics of magnetic orientation: strengthening the interface between theory and experimental design. J. R. Soc. Interface **7**, S179–S191 (2010). Jan
53. Le, T., Mayaram, K., Fiez, T.: Efficient far-field radio frequency energy harvesting for passively powered sensor networks. IEEE J. Solid-State Circuits **43**(5), 1287–1302 (2008)
54. Lahiri, I., Oh, S., Hwang, J.Y., Cho, S., Sun, Y., Banerjee, R., Choi, W.: High capacity and excellent stability of lithium ion battery anode using interface-controlled binder-free multiwall carbon nanotubes grown on copper. ACS Nano **4**(6), 3440–3446 (2010)
55. Malcolm, R.: A mechanism by which the hair cells of the inner ear transduce mechanical energy into a modulated train of action potentials. J. Gen. Physiol. **63**(6), 757 (1974)
56. Meyer, C.J., Alenghat, F.J., Rim, P., Fong, J.H., Fabry, B., Ingber, D.E.: Mechanical control of cyclic amp signalling and gene transcription through integrins. Nat. Cell Biol. **2**(9), 666–8 (2000)
57. Meirovitch, L.: Fundamentals of Vibrations (2001)
58. McSpadden, J.O., Mankins, J.C.: Space solar power programs and microwave wireless power transmission technology. IEEE Microw. Mag. **3**(4), 46–57 (2002)
59. Muxworthy, A.R., Williams, W.: Critical superparamagnetic/single-domain grain sizes in interacting magnetite particles: implications for magnetosome crystals. J. R. Soc. Interface **6**(41), 1207–12 (2009)
60. Mohan, A., Woo, G., Hiura, S., Smithwick, Q., Raskar, R.: Bokode: imperceptible visual tags for camera based interaction from a distance. ACM Trans. Graph. **28**(3):98:1–98:8 (2009)
61. Mishra, D., De, S., Jana, S., Basagni, S., Chowdhury, K., Heinzelman, W.: Smart RF energy harvesting communications: challenges and opportunities. IEEE Commun. Mag. **53**(4), 70–78 (2015)
62. The Near Field Communication Forum. http://www.nfc-forum.org/

63. Nelson, P.C., Radosavljević, M., Bromberg, S.: Biological Physics: Energy, Information, Life, p. 630, Jan 2008
64. Nakano, T., Suda, T., Koujin, T., Haraguchi, T., Hiraoka, Y.: Molecular communication through gap junction channels. Trans. Comput. Syst. Biol. X (2008)
65. Nakano, T., Suda, T., Moore, M., Egashira, R., Enomoto, A., Arima, K.: Molecular communication for nanomachines using intercellular calcium signaling. In: 2005 5th IEEE Conference on Nanotechnology, pp. 478– 481, vol. 2 (2005)
66. Phizicky, E.M., Fields, S.: Protein-protein interactions: methods for detection and analysis. Microbiol. Rev. **59**(1), 94–123 (1995)
67. Poon, A., O'Driscoll, S., Meng, T.: Optimal frequency for wireless power transmission into dispersive tissue. IEEE Trans. Antennas Propag. **58**(5), 1739–1750 (2010)
68. Poon, A.S.Y.: Miniaturization of implantable wireless power receiver. Conf. Proc. IEEE Eng. Med. Biol. Soc. **2009**, 3217–20 (2009)
69. PowerCast Corporation. http://www.powercastco.com/technology/powerharvester-receivers
70. Proakis, J.G., Salehi, M.: Digital Communications (2008)
71. Rabaey, J.: Wireless beyond the third generation-facing the energy challenge. In: Low Power Electronics and Design, Jan 2002
72. Rice, S.O.: Mathematical Analysis of Random Noise
73. RamRakhyani, A.K., Lazzi, G.: On the design of efficient multi-coil telemetry system for biomedical implants. IEEE Trans. Biomed. Circuits Syst. **PP**(99), 1 (2012)
74. Raghunathan, V., Schurgers, C., Park, S., Srivastava, M.B.: Energy-aware wireless microsensor networks. IEEE Signal Process. Mag. **19**(2), 40–50 (2002)
75. Sazonova, V.A.: A Tunable Carbon Nanotube Resonator (2006)
76. Soloveichik, D., Cook, M., Winfree, E., Bruck, J.: Computation with finite stochastic chemical reaction networks. Nat. Comput. Int. J. **7**(4) (2008)
77. Shetty, R.P., Endy, D., Knight, T.F.: Engineering biobrick vectors from biobrick parts. J. Biol. Eng. **2**, 5 (2008)
78. Sidles, J.A.: Spin microscopy's heritage, achievements, and prospects. Proc. Nat. Acad. Sci. USA **106**(8), 2477–8 (2009)
79. Stipe, B., Mamin, H., Stowe, T., Kenny, T., Rugar, D.: Magnetic dissipation and fluctuations in individual nanomagnets measured by ultrasensitive cantilever magnetometry. Phys. Rev. Lett. **86**(13), 2874–2877 (2001)
80. Shah, R.C., Rabaey, J.M.: Energy Aware Routing for Low Energy Ad Hoc Sensor Networks, vol. 1, pp. 350–355, Mar 2002
81. Schurgers, C., Raghunathan, V., Srivastava, M.B.: Power management for energy-aware communication systems. ACM Trans. Embed. Comput. Syst. **2**(3), 431–447 (2003)
82. Samanta, B., Yan, H., Fischer, N.O., Shi, J., Jerry, D.J., Rotello, V.M.: Protein-passivated fe3o4 nanoparticles: low toxicity and rapid heating for thermal therapy. J. Mater. Chem. **18**(11), 1204 (2008)
83. Sample, A.P., Yeager, D.J., Powledge, P.S., Mamishev, A.V., Smith, J.R.: Design of an rfid-based battery-free programmable sensing platform. IEEE Trans. Instrum. Measur. **57**(11), 2608–2615 (2008)
84. Takata, H., Kogure, O., Murase, K.: Matrix-addressed Liquid Crystal Displays, vol. 18, pp. 72 (1972)
85. Tanaka, K., Kenichiro, M., Takahashi, M., Ishii, T., Sasaki, S.: Development of bread board model for microwave power transmission experiment from space to ground using small scientific satellite. pp. 191–194, May 2012
86. Visser, H.J., Reniers, A.C.F., Theeuwes, J.A.C.: Ambient rf Energy Scavenging: Gsm and Wlan Power Density Measurements, pp. 721–724, Oct 2008
87. Wang, J.: Carbon-nanotube based electrochemical biosensors: a review. Electroanalysis (2005)
88. Weinstein, R.: RFID: a technical overview and its application to the enterprise. IT Prof. **7**(3), 27–33 (2005)
89. Winklhofer, M., Kirschvink, J.L.: A quantitative assessment of torque-transducer models for magnetoreception. J. R. Soc. Interface **7**, S273–S289 (2010)

90. Weiss, B.P., Kim, S.S., Kirschvink, J.L., Kopp, R.E., Sankaran, M., Kobayashi, A., Komeili, A.: Magnetic tests for magnetosome chains in martian meteorite alh84001. Proc. Nat. Acad. Sci. USA **101**(22), 8281–8284 (2004)
91. Wendt, T.M., Reindl, L.M.: Wake-Up Methods to Extend Battery Life Time of Wireless Sensor Nodes, pp. 1407–1412, May 2008
92. Yaghjian, A.: An overview of near-field antenna measurements. IEEE Trans. Antennas Propag. (1986)
93. Yoo, T.-W., Chang, K.: Theoretical and experimental development of 10 and 35 ghz rectennas. IEEE Trans. Microw. Theory Tech. **40**(6), 1259–1266 (1992)
94. Zahradka, K., Slade, D., Bailone, A., Sommer, S., Averbeck, D., Petranovic, M., Lindner, A.B., Radman, M.: Reassembly of shattered chromosomes in deinococcus radiodurans. Nature **443**(7111), 569–573 (2006)

Efficient Protocols for Peer-to-Peer Wireless Power Transfer and Energy-Aware Network Formation

Adelina Madhja, Sotiris Nikoletseas, Theofanis P. Raptis, Christoforos Raptopoulos and Dimitrios Tsolovos

Abstract Wireless power transfer provides the potential to efficiently replenish the energy and prolong the lifetime of nodes in ad hoc networks. Current state-of-the-art studies utilize strong charger stations (equipped with large batteries) with the main task of transmitting their available energy to the network nodes. Different to these works, in this chapter, we investigate interactive, "peer-to-peer" wireless energy exchange in populations of resource-limited mobile agents, without the use of any special chargers. The agents in this model are capable of mutual energy transfer, acting both as transmitters and receivers of wireless power. In such types of ad hoc networks, we propose protocols that address two important problems: the problem of energy balance between agents and the problem of distributively forming a certain network structure (a star) with an appropriate energy distribution among the agents. We evaluate key performance properties (and their trade-offs) of our protocols, such as their energy and time efficiency, as well as the achieved distance to the target energy distribution.

A. Madhja (✉) · S. Nikoletseas (✉) · C. Raptopoulos
Department of Computer Engineering and Informatics, University of Patras, Computer Technology Institute and Press "Diophantus" (CTI), Patras, Greece
e-mail: madia@ceid.upatras.gr

S. Nikoletseas
e-mail: nikole@cti.gr

C. Raptopoulos
e-mail: raptopox@ceid.upatras.gr

T. P. Raptis
Institute of Informatics and Telematics, National Research Council, Pisa, Italy
e-mail: theofanis.raptis@iit.cnr.it

D. Tsolovos
Inria Saclay-Ile-de-France & DAVID Lab, University of Versailles, Versailles, France
e-mail: dimitrios.tsolovos@inria.fr

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_14

1 Introduction

In many current application domains, such as medical and environmental monitoring, industrial automation, wireless sensor networks, intelligent transportation systems, etc., the need for battery-free ultralow-power devices, possibly wearable, or implantable, is increasing dramatically [7]. Recently, there has been an increasing interest to combine near-field communication capabilities and wireless power transfer in the same portable device, allowing mobile agents carrying the devices to wirelessly exchange energy. The near-field behavior of a pair of closely coupled transmitting and receiving dual-band printed monopole antennas (suitable for mobile phone applications) can make it possible to achieve both far-field performance and near-field power transfer efficiency (from 35 to 10%) for devices located few centimeters apart [9]. Further developments on the circuit design can render a device capable of achieving bi-directional, highly efficient wireless power transfer and both can be used as transmitter and as a receiver [34, 38]. In this context, energy harvesting and wireless power transfer capabilities are integrated, enabling each device to act on demand either as a wireless energy provider or as an energy harvester.

Populations of such devices have to operate under severe limitations in their computational power, data storage, the quality of communication and most crucially, their available amount of energy. For this reason, the efficient distributed cooperation of the agents towards achieving large computational and communication goals is a challenging task. An important goal in the design and efficient implementation of large networked systems is to save energy and keep the network functional for as long as possible [19, 37]. This can be achieved using wireless power transfer as an energy exchange enabling technology and applying interaction protocols among the agents which guarantee that the available energy in the network can be eventually distributed in a balanced way.

In this chapter, we present our recent line of research (presented in [20, 21, 27–29]) on a new model for configuring the wireless power transfer process in networked systems of mobile agents. Inspired by the Population Protocol model of [3] and [4], this is the first study (to the best of our knowledge) of bi-directional, interactive wireless charging in populations of mobile peers.

More specifically, we review two important problems under the model of interactive wireless power transfer:

- *Energy balance in the population*: We provide an upper bound on the time that is needed to reach energy balance in the population at the loss-less case, and we investigate the complex impact of the energy levels diversity in the lossy case; also, we highlight several key elements of the charging procedure. We provide three interaction protocols which take into account the different aspects of the charging procedure and achieve different performance trade-offs.
- *Energy-aware star network formation in the population*: For the first time, we *introduce energy issues* in network construction protocols. More specifically, the

network agents can exchange energy when they interact. In this chapter, we provide protocols that construct the *star* network structure and propose *a corresponding target energy distribution*.

2 Related Work

Wireless power transfer is an emerging technology which recently has been studied in various perspectives. Most of them use powerful entities called chargers, which are able to carry large amounts of energy and transfer it to the network devices. In [30, 31], the authors conducted experiments using real devices to evaluate their proposed protocols that maximize the energy efficiency and achieve energy balance. In [22, 23, 41], the authors proposed coordination protocols for multiple chargers. The minimum number of the required chargers is investigated in [8]. Both data gathering and energy transfer is studied in [17] and [46].

Numerous works suggest the employment of mobile wireless energy chargers in networks of sensor nodes, by combining energy transfer with data transmission and routing [12, 13, 45], or providing distributed and centralized solutions [42, 43]. In [18], the authors study the case of collaborative charging where chargers can charge each other, which in [23] was extended to a hierarchical structure. In [2], the authors investigated the case where the nodes are mobile, as in [30], but the charger is mobile too.

Other works focus on multi-hop energy transfer in stationary networks [35, 44], as well as UAV-assisted charging of ground sensors [10, 14]. Most of those wireless power transfer applications have also been verified experimentally, using real device prototypes [26, 33]. Although all those works provide nice solutions on the efficient charging of networks which is comprised of next- generation devices, none of them investigates the bi-directional charging procedure in populations of mobile agents.

Using bi-directional wireless power transfer technology [34], it is possible to both transmit and receive energy. The corresponding circuits can be embedded to the mobile agents in the population protocols a fundamental study of which is provided in [6]. In [24], the authors introduced the Arithmetic Population Protocol model where the agents with limited capabilities are able to compute order statistics of their arithmetic input values. More specifically, the agents have a fixed number of registers and the joint transition function allows them to only make comparisons and copy/paste operations on the values stored on their registers. In contrast, our model allows the agents to compare their registered values and to update them, according to the protocols. It is possible to construct a specific network structure as described in [25]. In [28], the authors proposed the model where the agents are able to perform peer-to-peer energy exchanges. We extend their model by letting the energy loss factor be different in each energy exchange.

Last but not least, the book [32], is the first systematic exposition on the domain of wireless energy transfer in ad hoc communication networks. It selectively spans a coherent, large spectrum of fundamental aspects of wireless energy transfer, such

as mobility management in the network, combined wireless energy and information transfer, energy flow among network devices, joint activities with wireless energy transfer (routing, data gathering and solar energy harvesting), and safety provisioning through electromagnetic radiation control, as well as fundamental and novel circuits and technologies enabling the wide application of wireless energy. In this work, we do not address the communication layer. For interesting discussions of said issues the reader may refer to [11, 40]. Also, the reader may refer to [39] for a detailed survey on data dissemination techniques in Delay-Tolerant Networks.

3 The Model

We consider a population of m mobile agents denoted as $M = \{u_1, u_2, \ldots, u_m\}$, each one equipped with a *battery cell*, a *wireless power transmitter*, and a *wireless power receiver*. Additionally, each agent u has a *state* from a set of states Q and a small *local memory* consisting of a small number of registers. For any time $t \geq 0$ and agent u, we denote by:

$$C_t(u) \stackrel{def}{=} (E_t(u), q_t(u), R_t(u)) \tag{1}$$

the *configuration of u at t*, where $E_t(u)$ (respectively $q_t(u)$, $R_t(u)$) is the *energy level* (respectively state and memory) of agent u at time t. The *relationship* between any pair of agents $\{u, u'\}$ is further characterized by a *connection state* from a set of states Q' (different from Q); here, we set $Q' = \{0, 1\}$. In particular, for any pair of agents u, u', if their connection state $q_t(u, u')$ at time t is equal to 1, then we say that u is connected to u' at t; otherwise (i.e., if $q_t(u, u') = 0$), they are disconnected.

Because of the limitations of current technology, the peers have to be within a few centimeters from each other, and the size of the overall network is in the order of a few meters. The movement of the agents does not follow any specific pattern, but whenever two agents meet (e.g., whenever their trajectory paths intersect or the agents come sufficiently close), they can interact according to an *interaction protocol* P; all agents run the *same* protocol P. In particular, whenever agents u, u' interact, they modify (a) their respective configurations (i.e., they exchange energy, modify their states and local memory) and (b) their connection state according to P. Formally, we assume that time is discrete and that if agents u, u' interact at time t, they communicate their configurations and current connection state and they jointly modify them as follows:

$$(C_{t+1}(u), C_{t+1}(u'), q_{\{u,u'\}}(t+1)) = P(C_t(u), C_t(u'), q_t(u, u')). \tag{2}$$

The configurations of all other agents, as well as every other connection state (including those involving agents u or u' with other agents) remain unchanged.

Due to the nature of wireless power technology (e.g., RF-to-DC conversion, materials and wiring used in the system, objects near the devices, etc.), any transfer of

energy induces *energy loss*. Therefore, whenever an agent u transfers energy ε to agent u', the amount of energy that the latter actually receives is $(1-\beta)\cdot\varepsilon$, where β is a parameter depending on the environment and the equipment for energy transfer available to the agents. In our experiments, we assume different values of β, which are *not known* by the agents. More specifically, we explore the behavior of our protocols in several cases where β is a constant value, as well as in the more general case, where the value of β can be *different* in every interaction. In particular, we assume that in every interaction, β is an *independent random variable* that follows some distribution (e.g., the normal distribution). For simplicity, we do not take into account energy loss due to movement or other activities of the agents explicitly, as this is besides the focus of our work. For a more realistic study of the physical properties of Wireless Power Transfer, we refer the reader to [15].

In fact, we assume that most devices can be carried by individuals or other moving entities that have their own agenda, and thus devices interact when the latter happens to come in close proximity. In the most general setting, interactions between agents are planned by a *scheduler* (that satisfies certain fairness conditions ensuring that all possible interactions will eventually occur), which can be used to abstract the movement of the agents. To allow for nontrivial results in our experimental evaluation of our algorithmic solutions, here we consider a widely accepted special case of fair scheduler, namely the *probabilistic scheduler*, which was introduced in [5]. According to the probabilistic scheduler, in every time step, a single interacting pair of agents is selected independently and uniformly at random among all $\binom{m}{2}$ pairs of agents in the population.

A crucial assumption of this model (which is inspired by the population protocols model [5] and network constructors [25]) is that agents do not share memory or exchange messages unless they interact. Furthermore, agents are computationally weak machines that cannot grasp the full structure and status of the entire population. Nevertheless, through pair-wise interactions, agents are required to collect eventually *converge* to a stable state.

4 Problem Definition and Metrics

In [25], the authors define the *population network* at time t to be the simple, undirected graph G_t with vertex set as the set of agents M and edge set the set of pairs of agents u, u' that have $q_{\{u,u'\}} = 1$. In particular, they design protocols (which they call *network constructors*) that eventually converge to certain graph structures. We significantly generalize the definition of network constructors to take into account the energy levels of the agents in the population; we call this *energy-aware network formation*. To this end, we use two metrics: the *structural distance* and the *energy distance*.

Formally, let H be a *target graph* on m vertices. For two graphs H, G on the same vertex set M, we denote by $H \triangle G$ the hamming distance between those graphs, i.e.,

$$H \triangle G \stackrel{def}{=} \sum_{e} |\mathbf{1}_e(H) - \mathbf{1}_e(G)|, \tag{3}$$

where the summation is overall $\binom{m}{2}$ possible edges and $\mathbf{1}_e(H)$ (respectively $\mathbf{1}_e(G)$) is the indicator variable for the existence of e in H (respectivley G). We define the *structural distance* of the population from the target graph H at time t as follows:

$$\delta_t^s(H, G_t) \stackrel{def}{=} \min_{G \sim G_t} H \triangle G, \tag{4}$$

where G_t is the population network at time t and the minimum is taken overall graphs G that are isomorphic to G_t.

The energy distance is defined in analogy to the *total variation distance* in probability theory and stochastic processes [1, 16]. Let $\mathscr{E}^*$ be a *target distribution*, defined on $[m] = \{1, 2, \ldots, m\}$ and, for any $t \geq 0$, let $\mathscr{E}_t$ be the relative energy distribution at time t given by $\mathscr{E}_t(u) = \frac{E_t(u)}{\sum_u E(u)}$, $u \in M$. Let also $\Sigma(m)$ be the set of permutations of $[m]$. We define the *energy distance* of the population from the target energy distribution $\mathscr{E}^*$ at time t as follows:

$$\delta_t^e(\mathscr{E}^*, \mathscr{E}_t) \stackrel{def}{=} \min_{\sigma \in \Sigma(m)} \frac{1}{2} \sum_{i=1}^{m} |\mathscr{E}_i^* - \mathscr{E}_t(\sigma(u_i))|, \tag{5}$$

where the minimum is among all permutations of $[m]$, $\mathscr{E}_t(\sigma(u_i))$ is the relative energy level of agent $\sigma(u_i)$ at time t and $\mathscr{E}_i^*$ is the target distribution at point i of its domain.

The general formulation of the problem that we consider in this chapter is as follows:

Definition 1 *(Energy-aware network formation)* Consider a population M of agents. Let H be a target graph on M and $\mathscr{E}^*$ a target distribution. Let also ε be a small positive constant. Assuming the probabilistic scheduler, find a protocol that, when run by the agents in the population, there is $t \geq 0$ such that:

1. $\delta_t^s(H, G_t) = 0$,
2. $\delta_t^e(\mathscr{E}^*, \mathscr{E}_t) \leq \varepsilon$ and
3. the total energy loss is minimized, i.e., $E_0(M) - E_t(M) = \sum_u E_0(u) - \sum_u E_t(u)$ is as small as possible.

In this work, we consider two special cases of the Energy-Aware Network Formation Problem: (a) the *Population Energy Balance* problem and (b) the *Energy-Aware Star Formation* problem. In the Population Energy Balance Problem, the structure is irrelevant (alternatively, this version of the problem may be thought as the energy-aware construction of the complete graph), and the ultimate goal is to achieve approximate energy balance at the minimum energy loss across agents in M. For this special case, we also make the additional assumption that the energy loss factor is constant for each interaction (rather than a random variable). In the second problem, we consider the construction of one of the most basic graph structures, namely the star.

Furthermore, we assume that the target distribution is such that *the relative energy level of each node is proportional to its degree*. Our motivation for this problem comes from the fact that star formations usually arise in wireless networks when nodes are organized in a cluster, in which case a cluster head is selected to which all communication is forwarded. In view of this, the energy level of the cluster head should be proportional to the number of nodes in its cluster. Therefore, the target energy level of the central node of the star at time t should be $a = \frac{E_t(M)}{2}$, while the target energy level of a peripheral node should be $b = \frac{E_t(M)}{2(m-1)}$. Setting without loss of generality $\mathscr{E}^* = \{a, b, b, \ldots, b\}$, the minimum of $\frac{1}{2}\sum_{i=1}^{m} |\mathscr{E}_i^* - \mathscr{E}_t(\sigma(u_i))|$ (which is equal to the energy distance $\delta_t^e(\mathscr{E}^*, \mathscr{E}_t)$) is attained by choosing any permutation σ that assigns the agent with the largest energy level to u_1.

5 The Population Energy Balance Problem

Let U be the uniform distribution on M. We will say that the population has energy balance ε at time t if and only if $\delta_t^e(U, \mathscr{E}_t) \leq \varepsilon$. It is evident from the definition of our model that $\delta_t^e(U, \mathscr{E}_t)$ is a random variable, depending on the specific distribution of energies in the population and the choice is made by the probabilistic scheduler at time t. Therefore, we are rather interested in *protocols that reduce the total variation distance on expectation with the smallest energy loss*. Furthermore, we measure the efficiency of a protocol P by the expected energy loss and the expected time needed for the protocol to reach energy balance.

5.1 Loss-Less Energy Transfer

In this section, we present a very simple protocol for energy balance in the case of loss-less energy transfer, i.e., for $\beta = 0$ (Protocol 1). The protocol basically states that, whenever two agents u, u' interact, they split their cumulative energy in half.

Protocol 1: Oblivious-Share P_{OS}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$

1 $P_{\mathrm{OS}}(\varepsilon_u, \varepsilon_{u'}) = \left(\frac{\varepsilon_u+\varepsilon_{u'}}{2}, \frac{\varepsilon_u+\varepsilon_{u'}}{2}\right)$.

In the following Lemma, we show that, when all agents in the population use protocol P_{OS}, the total variation distance decreases in expectation. The proof not only

leads to an upper bound on the time needed to reach energy balance (see Theorem 1), but more importantly, highlights several key elements of the energy transfer process, which we exploit when designing interaction protocols for the case $\beta > 0$ in Sect. 5.2.

Lemma 1 *Let M be a population of chargers using protocol P_{OS}. Assuming interactions are planned by the probabilistic scheduler and there is no loss from energy exchanges, we have that*

$$\mathbb{E}[\delta_t^e(\mathscr{E}_t, U)|\mathscr{E}_{t-1}] \leq \left(1 - \frac{2}{\binom{m}{2}}\right) \delta_t^e(\mathscr{E}_{t-1}, U). \tag{6}$$

Proof We first note that, since we are in the loss-less case, i.e., $L(\varepsilon) = 0$, for any transfer of an amount ε of energy, we have that $E_t(M) = E_0(M)$, for any t (i.e., the total energy amount remains the same). Furthermore, $U(x) = \frac{1}{m}$, for all $x \in M$.

Define $\Delta_t \stackrel{def}{=} \delta_t^e(\mathscr{E}_t, U) - \delta_t^e(\mathscr{E}_{t-1}, U)$ and assume that, at time t, agents u, u' interact. By a simple observation, since the state of every other agent remains the same, we have that

$$\Delta_t = \left|\frac{\mathscr{E}_{t-1}(u) + \mathscr{E}_{t-1}(u')}{2} - \frac{1}{m}\right| - \frac{1}{2}\left(\left|\mathscr{E}_{t-1}(u) - \frac{1}{m}\right| + \left|\mathscr{E}_{t-1}(u') - \frac{1}{m}\right|\right) \tag{7}$$

For any charger $x \in M$ and time $t \geq 0$, set now $z_t(x) \stackrel{def}{=} \mathscr{E}_t(x) - \frac{1}{m}$. Let also $A_t^+ \subseteq M$ (respectivley A_t^-, $A_t^=$) be the set of chargers such that $z_t(x)$ is positive (respectivley negative and equal to 0).

By direct computation using Eq. (7), we can see that the total variation distance at time t decreases (i.e., Δ_t is strictly less than 0) if and only if $u \in A_t^+$, $u' \in A_t^-$, or $u \in A_t^-$, $u' \in A_t^+$; otherwise it remains the same (i.e., $\Delta_t = 0$). Indeed, using the numbers $z_{t-1}(x)$, $x \in M$, for the sake of compactness, we distinguish the following cases:

Case I: If $z_{t-1}(u)z_{t-1}(u') \geq 0$, then $\Delta_t = 0$.
Case II: If $z_{t-1}(u)z_{t-1}(u') < 0$ and $|z_{t-1}(u)| \geq |z_{t-1}(u')|$, then $\Delta_t = -2|z_{t-1}(u')|$.
Case III: If $z_{t-1}(u)z_{t-1}(u') < 0$ and $|z_{t-1}(u)| < |z_{t-1}(u')|$, then $\Delta_t = -2|z_{t-1}(u)|$.

Since interactions are planned by the probabilistic scheduler, i.e., any specific pair u, u' of agents is chosen for interaction at time t with probability $\frac{1}{\binom{m}{2}}$, by linearity of expectation and Eq. (7), we get:

$$\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] = -\frac{1}{\binom{m}{2}}\left(\sum_{x\in M} 2|z_{t-1}(x)| \cdot |\{y : |z_{t-1}(y)| \geq |z_{t-1}(x)|, z_{t-1}(x)z_{t-1}(y) < 0\}|\right.$$
$$\left. - \sum_{x\in M} |z_{t-1}(x)| \cdot |\{y : |z_{t-1}(y)| = |z_{t-1}(x)|, z_{t-1}(x)z_{t-1}(y) < 0\}|\right), \tag{8}$$

where we subtracted

$$\sum_{x \in M} |z_{t-1}(x)| \cdot |\{y : |z_{t-1}(y)| = |z_{t-1}(x)|, z_{t-1}(x)z_{t-1}(y) < 0\}| \tag{9}$$

from the above sum, since the contribution $-2|z_{t-1}(x)|$ of agent x is counted twice for agents x, y such that $|z_{t-1}(x)| = |z_{t-1}(y)|$ (once for x and once for y). Notice also that, in the above sum, we can ignore agents $x \in A^{=}_{t-1}$, since their contribution to $\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}]$ is 0.

In order to give a formula for $\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}]$ that is easier to handle, consider a complete ordering σ_{t-1} of the agents $x \in A^{+}_{t-1} \cup A^{-}_{t-1}$ in increasing value of $|z_{t-1}(x)|$, breaking ties arbitrarily. We will write $x <_{\sigma_{t-1}} y$ if agent x is "to the left" of agent y in σ_{t-1}, or equivalently $\sigma_{t-1}(x) < \sigma_{t-1}(y)$. We can then see that, the contribution of an agent $x \in A^{+}_{t-1}$ (respectively $x \in A^{-}_{t-1}$) to $\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}]$ is $|z_{t-1}(x)|$ multiplied by the number of agents in A^{-}_{t-1} (respectivley A^{+}_{t-1}) that are "to the right" of x in σ_{t-1} (i.e., agents y such that $x <_{\sigma_{t-1}} y$, for which $z_{t-1}(x)z_{t-1}(y) < 0$). Therefore, Eq. (8) becomes

$$\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] = -\frac{2}{\binom{m}{2}} \sum_{x \in M} |z_{t-1}(x)| \cdot |\{y : x <_{\sigma_{t-1}} y, z_{t-1}(x)z_{t-1}(y) < 0\}|. \tag{10}$$

Assume now, without loss of generality, that the "rightmost" agent in σ_{t-1} is some $y^* \in A^{+}_{t-1}$. By the above equation, we then have that the contribution $-|z_{t-1}(x)|$ of every agent $x \in A^{-}_{t-1}$ is counted at least once (because of y^*, since $x <_{\sigma_{t-1}} y^*$ and $z_{t-1}(x)z_{t-1}(y^*) < 0$). Therefore,

$$\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] \leq -\frac{2}{\binom{m}{2}} \sum_{x \in A^{-}_{t-1}} |z_{t-1}(x)|. \tag{11}$$

But using a standard result on total variation distance (see for example, [1]), we have that

$$\delta^e_t(\mathscr{E}_{t-1}, U) = \sum_{x \in A^{-}_{t-1}} |z_{t-1}(x)| = \sum_{y \in A^{+}_{t-1}} |z_{t-1}(y)| \tag{12}$$

Therefore, we have that

$$\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] \leq -\frac{2}{\binom{m}{2}} \delta^e_t(\mathscr{E}_{t-1}, U) \tag{13}$$

which completes the proof. □

It is worth noting that the upper bound of Lemma 1 is tight when the distribution of energies is such that there is only one agent with energy above or below the average.

We now use Lemma 1 to prove that protocol P_{OS} is quite fast in achieving energy balance in the loss-less case.

Theorem 1 *Let M be a population of chargers using protocol P_{OS}. Let also $\tau_0(c)$ be the time after which $\mathbb{E}[\delta_t^e(\mathscr{E}_{\tau_0(c)}, U_{\tau_0(c)})] \leq c$, assuming interactions are planned by the probabilistic scheduler and there is no loss from energy exchanges. Then $\tau_0(c) \leq \frac{1}{2}\binom{m}{2} \ln\left(\frac{\delta_t^e(\mathscr{E}_0,U)}{c}\right)$, where $\delta_t^e(\mathscr{E}_0, U)$ is the total variation distance between the initial energy distribution and the uniform energy distribution.*

Proof Taking expectations in the upper bound inequality from Lemma 1, we have that

$$\mathbb{E}[\mathbb{E}[\delta_t^e(\mathscr{E}_t, U)|\mathscr{E}_{t-1}]] \leq \left(1 - \frac{2}{\binom{m}{2}}\right) \mathbb{E}[\delta_t^e(\mathscr{E}_{t-1}, U)] \tag{14}$$

or equivalently

$$\mathbb{E}[\delta_t^e(\mathscr{E}_t, U)] \leq \left(1 - \frac{2}{\binom{m}{2}}\right) \mathbb{E}[\delta_t^e(\mathscr{E}_{t-1}, U)]. \tag{15}$$

Iterating the above inequality, we then have that

$$\mathbb{E}[\delta_t^e(\mathscr{E}_t, U)] \leq \left(1 - \frac{2}{\binom{m}{2}}\right)^t \delta_t^e(\mathscr{E}_0, U) \leq e^{-\frac{2t}{\binom{m}{2}}} \delta_t^e(\mathscr{E}_0, U). \tag{16}$$

Consequently, for any $t \geq \frac{1}{2}\binom{m}{2} \ln\left(\frac{\delta_t^e(\mathscr{E}_0,U_0)}{c}\right)$, we have that $\mathbb{E}[\delta_t^e(\mathscr{E}_t, U_t)] \leq c$, which concludes the proof. □

5.2 Energy Transfer with Loss

In this section, we consider the more natural case where every transfer of energy ε induces energy loss $L(\varepsilon) = \beta\varepsilon$, for some $0 < \beta < 1$. The main technical difficulty that arises in this case when considering the total variation distance change $\Delta_t = \delta_t^e(\mathscr{E}_t, U) - \delta_t^e(\mathscr{E}_{t-1}, U)$ is that any energy transfer between agents u and u' affects also the relative distance of energy levels of noninteracting agents from the total average. More precisely, after u, u' exchange energy ε at time t, we have $E_t(M) = E_{t-1}(M) - \beta\varepsilon$. Therefore, for any noninteracting agent x at time t, we have

$$|z_t(x)| \stackrel{def}{=} \left|\mathscr{E}_t(x) - \frac{1}{m}\right| = \left|\frac{E_t(x)}{E_t(M)} - \frac{1}{m}\right| \neq \left|\frac{E_{t-1}(x)}{E_{t-1}(M)} - \frac{1}{m}\right| \stackrel{def}{=} |z_{t-1}(x)| \tag{17}$$

As a consequence, straightforward generalizations of simple protocols like P_{OS} do not perform up to par in this case.

In particular, there are specific worst-case distributions of energies for which the total variation distance increases on expectation after any significant energy exchange. As a fictitious example, consider a population of m agents, for some

Protocol 2: Small-Transfer P_{ST}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$
if $\varepsilon_u \geq \varepsilon_{u'} - d\varepsilon$ **then**
  $P_{\mathrm{ST}}(\varepsilon_u, \varepsilon_{u'}) = (\varepsilon_u - d\varepsilon, \varepsilon_{u'} + (1-\beta)d\varepsilon)$
else if $\varepsilon_{u'} \geq \varepsilon_u - d\varepsilon$ **then**
  $P_{\mathrm{ST}}(\varepsilon_u, \varepsilon_{u'}) = (\varepsilon_u + (1-\beta)d\varepsilon, \varepsilon_{u'} - d\varepsilon)$
else if $|\varepsilon_u - \varepsilon_{u'}| < d\varepsilon$ **then**
  do nothing.

$m > \frac{2+\beta}{\beta}$. Furthermore, suppose that agents u_i, for $i \in [m-1]$ have energy m, while agent u_m has 0 energy at his disposal. The total variation distance in this example is $\frac{1}{m}$. Without loss of generality, consider now a variation of P_{OS} (adapted from the original version for $\beta = 0$ to the case of $\beta > 0$) according to which, whenever two agents u, u' interact, the agent with the largest amount of energy transfers $\frac{|\varepsilon_u - \varepsilon_{u'}|}{2}$ energy to the other.[1] We now have that after any significant energy exchange step, i.e., an interaction of u_m with any other agent, say x, according to protocol P_{OS}, the new energy level of u_m becomes $\frac{m}{2}(1-\beta)$, the new energy level of x becomes $\frac{m}{2}$, while the energy levels of all other agents remain the same. Therefore, the new total energy in the population becomes $m^2 - m - \frac{\beta}{2}m$, and the new total variation distance becomes[2]

$$\left(\frac{1}{m} - \frac{1}{2m-2-\beta}\right) + \left(\frac{1}{m} - \frac{1}{2m-2-\beta}(1-\beta)\right) = \frac{2}{m} - \frac{2-\beta}{2m-2-\beta} \quad (18)$$

which is strictly larger than $\frac{1}{m}$, for any $m > \frac{2+\beta}{\beta}$. We conclude that, in this example, the total variation distance increases also in expectation, as any interaction between pairs of agents that do not contain u_m does not change the energy distribution.

It is worth noting that, even though the above example is fictitious, our experiments verify our intuition that P_{OS} is not very suitable for energy balance in the case of lossy energy transfer. In particular, it seems that the energy lost with every step does not contribute sufficiently to the reduction of total variation distance between the distribution of energies and the uniform distribution. Our first attempt to overcome this problem was to only allow energy transfers between agents whose energy levels differ significantly. However, this did not solve the problem either (see our experimental results in Fig. 1), mainly because of interactions between agents that are both below the average energy. As our main contribution, we present in Sects. 5.2.1 and 5.2.2 two interaction protocols that seem to make the most of the energy lost in every step.

[1] Nevertheless, it is not hard to see that other variations of P_{OS} have similar problems.

[2] The total variation distance consists of two terms, since there are only two agents with energy levels below the average.

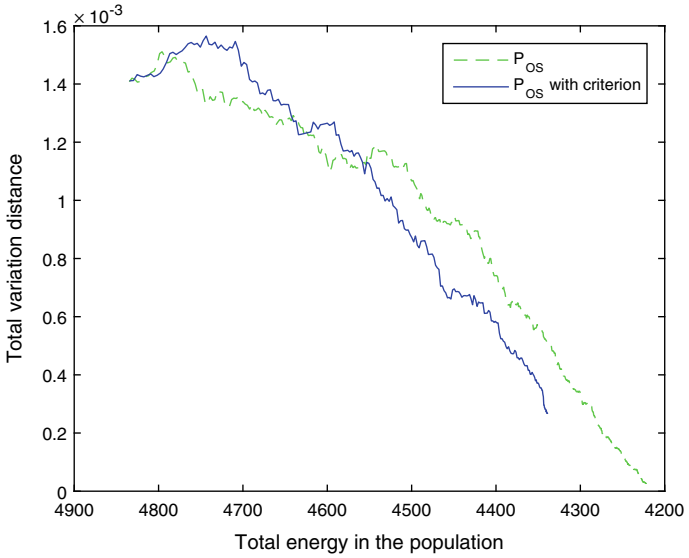


Fig. 1 Efficiency of the P_{OS} protocol applying the criterion that allows transfers between agents with significantly different energy levels, compared to the standard protocol version

5.2.1 Small-Transfer P_{ST}

In this subsection, we present the protocol Small-Transfer P_{ST} (Protocol 2), which suggests having only small energy transfers between interacting agents. Ideally, we only allow exchanges of infinitesimal energy $d\varepsilon$, which simplifies our analysis (in the experiments, we just choose a very small fixed value ε). Even though this idea is wasteful on time, we provide both analytic and experimental evidence that it achieves energy balance without wasting too much energy. For this very protocol, we realize the fact that under current technology, repetitive transfers of small energy amounts may be inefficient in view of the overhead incurred in any transfer but we expect future advances in wireless power transfer technology will eliminate this problem.

We prove the following lemma concerning the total variation distance change in a population of agents that use protocol P_{ST}.

Lemma 2 *Let M be a population of chargers using protocol P_{ST}. Given any distribution of energy $\mathscr{E}_{t-1}$, let $|A^+_{t-1}|$ (respectively $|A^-_{t-1}|$) be the number of agents with available energy above (respectively below) the current average. Assuming interactions are planned by the probabilistic scheduler, we have that*

$$\mathbb{E}[\Delta_t | \mathscr{E}_{t-1}] \leq \frac{4}{E_t(M)} \left(\beta - \frac{|A^+_{t-1}| \cdot |A^-_{t-1}|}{m(m-1)} \right). \tag{19}$$

Proof We will use the notation from Lemma 1. Let $a^+ = |A^+_{t-1}|$, $a^- = |A^-_{t-1}|$ and $a^= = |A^=_{t-1}|$. Assume without loss of generality, that at time t, the probabilistic scheduler selects agents u, u', such that $E_{t-1}(u) > E_{t-1}(u') - d\varepsilon$. Therefore, according to P_{ST}, agent u transfers energy $d\varepsilon$ to u', and so $E_t(u) = E_{t-1}(u) - d\varepsilon$ and $E_t(u') = E_{t-1}(u') + (1-\beta)d\varepsilon$. The energy level of every other charger remains unchanged. Furthermore, the new total energy in the population is:

$$E_t(M) = E_{t-1}(M) - \beta d\varepsilon. \tag{20}$$

A crucial observation for the analysis is that since P_{ST} only allows transfers of infinitesimal amounts of energy, after any useful interaction (i.e., interactions that change the distribution of energy in the population), the only agents that can potentially change the relative position of their energy levels to the average energy are those in $A^=_{t-1}$.

We now distinguish the following cases:

Case I: For any $x \in (A^+_{t-1} \cup A^=_{t-1})\backslash\{u, u'\}$, we have that $z_{t-1}(x) \geq 0$, so

$$z_t(x) = \frac{E_{t-1}(x)}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} > 0 \tag{21}$$

Therefore,

$$|z_t(x)| = \frac{E_{t-1}(x)}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} = \frac{E_{t-1}(M)}{E_t(M)}|z_{t-1}(x)| + \beta\frac{1}{m}\frac{1}{E_t(M)}d\varepsilon. \tag{22}$$

Case II: For any $x \in A^-_{t-1}\backslash\{u, u'\}$, we have that $z_{t-1}(x) < 0$, so

$$z_t(x) = \frac{E_{t-1}(x)}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} < 0 \tag{23}$$

for any infinitesimal energy transfer $d\varepsilon$. Therefore,

$$|z_t(x)| = \frac{E_{t-1}(x)}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} = \frac{E_{t-1}(M)}{E_t(M)}|z_{t-1}(x)| - \beta\frac{1}{m}\frac{1}{E_t(M)}d\varepsilon. \tag{24}$$

Case III: If $u \in A^-_{t-1} \cup A^=_{t-1}$, then $z_{t-1}(u) \leq 0$, so

$$z_t(u) = \frac{E_{t-1}(x) - d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} < 0 \tag{25}$$

since $E_{t-1}(u) \leq E_{t-1}(M)$ and $\beta \in (0, 1)$.

Furthermore, by assumption, $z_{t-1}(u') < 0$, and also (by the conditions of P_{ST}),

$$z_t(u') = \frac{E_{t-1}(u') + (1-\beta)d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} \leq \frac{E_{t-1}(x) - d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} < 0 \tag{26}$$

Therefore,

$$\begin{aligned} |z_t(u)| + |z_t(u')| &= \frac{1}{m} - \frac{E_{t-1}(x) - d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} + \frac{1}{m} - \frac{E_{t-1}(x) + (1-\beta)d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} \\ &= \frac{E_{t-1}(M)}{E_t(M)}(|z_{t-1}(u)| + |z_{t-1}(u')|) + \beta\left(1 - \frac{2}{m}\right)\frac{1}{E_t(M)}d\varepsilon. \end{aligned} \tag{27}$$

Case IV: If $u' \in A^+_{t-1} \cup A^=_{t-1}$, then $z_{t-1}(u') \geq 0$, so

$$z_t(u') = \frac{E_{t-1}(u') + (1-\beta)d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} \geq 0 \tag{28}$$

Furthermore, by assumption, $z_{t-1}(u) > 0$, and also (by the conditions of P_{ST}),

$$z_t(u) = \frac{E_{t-1}(x) - d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} > \frac{E_{t-1}(u') + (1-\beta)d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} \geq 0 \tag{29}$$

Therefore,

$$\begin{aligned} |z_t(u)| + |z_t(u')| &= \frac{E_{t-1}(x) - d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} + \frac{E_{t-1}(x) + (1-\beta)d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} \\ &= \frac{E_{t-1}(M)}{E_t(M)}(|z_{t-1}(u)| + |z_{t-1}(u')|) - \beta\left(1 - \frac{2}{m}\right)\frac{1}{E_t(M)}d\varepsilon. \end{aligned} \tag{30}$$

Case V: If $u \in A^+_{t-1}$ and $u' \in A^-_{t-1}$, then, similarly to the other cases we have $|z_{t-1}(u)| > 0$, $|z_t(u)| > 0$, $|z_{t-1}(u')| < 0$ and $|z_t(u')| < 0$. Therefore,

Case VI: If $u, u' \in A^=_{t-1}$ there is no change in the energy distribution.

$$\begin{aligned} |z_t(u)| + |z_t(u')| &= \frac{E_{t-1}(x) - d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} - \frac{1}{m} + \frac{1}{m} - \frac{E_{t-1}(x) + (1-\beta)d\varepsilon}{E_{t-1}(M) - \beta d\varepsilon} \\ &= \frac{E_{t-1}(M)}{E_t(M)}(|z_{t-1}(u)| + |z_{t-1}(u')|) - (2-\beta)\frac{1}{E_t(M)}d\varepsilon. \end{aligned} \tag{31}$$

Furthermore, the probability that the agents u, u', that are chosen for interaction by the probabilistic scheduler, are such that the conditions of case III (respectively case IV and case V) are satisfied, is $p_{\mathrm{III}} = \frac{a^-(a^- - 1) + 2a^-a^=}{m(m-1)}$ (respectively $p_{\mathrm{IV}} = \frac{a^+(a^+ - 1) + 2a^+a^=}{m(m-1)}$ and $p_{\mathrm{V}} = \frac{2a^-a^+}{m(m-1)}$). Putting it all together, by linearity of expectation, we have

$$
\begin{aligned}
&\mathbb{E}[\delta_t^e(\mathscr{E}_t, U)|\mathscr{E}_{t-1}] = \\
&= p_{\mathrm{III}} \cdot \left(\frac{E_{t-1}(M)}{E_t(M)} \delta_t^e(\mathscr{E}_{t-1}, U) - \frac{1}{E_t(M)} \left(-\beta - \frac{\beta(a^+ + a^= - a^-)}{m} \right) d\varepsilon \right) \\
&\quad + p_{\mathrm{IV}} \cdot \left(\frac{E_{t-1}(M)}{E_t(M)} \delta_t^e(\mathscr{E}_{t-1}, U) - \frac{1}{E_t(M)} \left(\beta - \frac{\beta(a^+ + a^= - a^-)}{m} \right) d\varepsilon \right) \\
&\quad + p_{\mathrm{V}} \cdot \left(\frac{E_{t-1}(M)}{E_t(M)} \delta_t^e(\mathscr{E}_{t-1}, U) - \frac{1}{E_t(M)} \left(2 - \beta - \frac{\beta(a^+ + a^= - a^-)}{m} \right) d\varepsilon \right) \\
&= (p_{\mathrm{III}} + p_{\mathrm{IV}} + p_{\mathrm{V}}) \frac{E_{t-1}(M)}{E_t(M)} \delta_t^e(\mathscr{E}_{t-1}, U) \\
&\quad + (p_{\mathrm{III}} + p_{\mathrm{IV}} + p_{\mathrm{V}}) \frac{1}{E_t(M)} \frac{\beta(a^+ + a^= - a^-)}{m} d\varepsilon \\
&\quad + \frac{\beta(p_{\mathrm{III}} + p_{\mathrm{V}} - p_{\mathrm{IV}}) - 2p_{\mathrm{V}}}{E_t(M)} d\varepsilon. \qquad (32)
\end{aligned}
$$

By rearranging, we have

$$
\begin{aligned}
\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] = &\left((p_{\mathrm{III}} + p_{\mathrm{IV}} + p_{\mathrm{V}}) \left(1 + \frac{\beta d\varepsilon}{E_t(M)} \right) - 1 \right) \delta_t^e(\mathscr{E}_{t-1}, U) + \\
&+ (p_{\mathrm{III}} + p_{\mathrm{IV}} + p_{\mathrm{V}}) \frac{1}{E_t(M)} \frac{\beta(a^+ + a^= - a^-)}{m} d\varepsilon + \\
&+ \frac{\beta(p_{\mathrm{III}} + p_{\mathrm{V}} - p_{\mathrm{IV}}) - 2p_{\mathrm{V}}}{E_t(M)} d\varepsilon \qquad (33)
\end{aligned}
$$

By now using the fact that $p_{\mathrm{III}}, p_{\mathrm{IV}}, p_{\mathrm{V}} \in [0, 1]$, $p_{\mathrm{III}} + p_{\mathrm{IV}} + p_{\mathrm{V}} \leq 1$ and the fact that the total variation distance between any two distributions is at most 1 (see e.g., [1]), we get

$$
\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] \leq \frac{4}{E_t(M)} \left(\beta - \frac{a^+ a^-}{m(m-1)} \right) \qquad (34)
$$

which completes the proof of the Lemma. □

It is worth noting that the upper bound on the total variation distance change from the above Lemma is quite crude (and can be positive if β is not small enough). However, this is mainly a consequence of our analysis; in typical situations, the upper bound that we get from inequality (34) can be much smaller. For example, if the energy distribution $\mathscr{E}_{t-1}$ at $t-1$ is such that $|A_{t-1}^+| \approx |A_{t-1}^-| \approx \frac{m}{2}$, (34) gives the bound $\mathbb{E}[\Delta_t|\mathscr{E}_{t-1}] \leq -\frac{1-\beta}{E_t(M)} d\varepsilon$, which is negative for any $\beta \in (0, 1)$. This is also verified by our experimental evaluation of P_{ST}. Nevertheless, the upper bound that we get from Lemma 2 highlights key characteristics of the interactive energy transfer process as we pass from loss-less (i.e., $\beta = 0$) to lossy energy transfer (i.e., $\beta > 0$).

Protocol 3: Online-Average P_{OA}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$

$\mathrm{avg}(u) = \frac{\mathrm{avg}(u)\cdot\mathrm{num}(u)+\varepsilon_{u'}}{\mathrm{num}(u)+1}$

$\mathrm{avg}(u') = \frac{\mathrm{avg}(u')\cdot\mathrm{num}(u')+\varepsilon_{u}}{\mathrm{num}(u')+1}$

$\mathrm{num}(u) = \mathrm{num}(u) + 1$

$\mathrm{num}(u') = \mathrm{num}(u') + 1.$

if $(\varepsilon_u > \mathrm{avg}(u)$ *and* $\varepsilon'_u \leq \mathrm{avg}(u'))$ *OR* $(\varepsilon_u \leq \mathrm{avg}(u)$ *and* $\varepsilon'_u > \mathrm{avg}(u'))$ **then**

  if $\varepsilon_u > \varepsilon_{u'}$ **then**

    $P_{\mathrm{OA}}(\varepsilon_u, \varepsilon_{u'}) = \left(\frac{\varepsilon_u+\varepsilon_{u'}}{2}, \frac{\varepsilon_u+\varepsilon_{u'}}{2} - \beta\frac{\varepsilon_u-\varepsilon_{u'}}{2}\right)$

  else if $\varepsilon_u \leq \varepsilon_{u'}$ **then**

    $P_{\mathrm{OA}}(\varepsilon_u, \varepsilon_{u'}) = \left(\frac{\varepsilon_u+\varepsilon_{u'}}{2} - \beta\frac{\varepsilon_{u'}-\varepsilon_u}{2}, \frac{\varepsilon_u+\varepsilon_{u'}}{2}\right)$

else

  do nothing.

5.2.2 Online-Average P_{OA}

By the analysis of the expected total variation distance change in Lemma 1 for energy transfer without losses, we can see that the total variation distance decreases when the interacting agents have energy levels that are on different sides of the average energy in the population. Using the notation form the proof of Lemma 1, if agents u, u' interact at time t, then we must either have $u \in A^+_{t-1}$ and $u' \in A^-_{t-1}$, or $u \in A^-_{t-1}$ and $u' \in A^+_{t-1}$, in order for the total variation distance $\delta^e_t(\mathscr{E}_t, U)$ to drop below $\delta^e_t(\mathscr{E}_{t-1}, U)$. The situation becomes more complicated when there are losses in energy transfers, but the analysis in Sect. 5.2.1 suggests that, under certain constraints on the energy distribution and the energy loss factor β, the total variation distance decreases whenever there is an interaction between a high relative energy agent and a low relative energy agent.

In view of the above, an ideal interaction protocol would only allow energy transfers between agents with energy levels that are on opposite sides of the average energy in the population. In particular, this would imply that, at any time t, each agent x would need to know the sign of $z_t(x) = \frac{E_t(x)}{E_t(M)} - \frac{1}{m}$, which is possible if x knows (in addition to its own energy level $E_t(x)$) the average energy $\frac{E_t(M)}{m}$ in the population. However, this kind of global knowledge is too powerful in our distributed model, since we assume that agents are independent and identical with each other. In particular, this implies that not only are agents not aware of other agents they have not yet interacted with, but also, that they have no way of knowing whether they have met with another agent at some point in the past.

The main idea behind our interaction protocol P_{OA} (Protocol 3) is that, even in our weak model of local interactions, agents are still able to compute local estimates of the average energy based on the energy levels of agents they interact with. To do this, every agent needs to keep track of the total number of interactions she has done,

as well as her current estimation for the average energy. This is accomplished by having each agent $x \in M$ maintaining two local registers, namely:

1. num(x), which is used to count the number of interactions that x has been involved in.
2. avg(x), which stores the current estimation of x for the average energy.

Furthermore, num(x) is initialized to 1, and avg(x) is initialized to $E_0(x)$. We give the formal description of our protocol below.

It is worth noting that P_{OA} may not perform up to par in the general case where interactions are planned by a potentially adversarial scheduler because the local estimates kept by agents for the average can be highly biased. On the other hand, in our experimental evaluation, we show that P_{OA} outperforms both P_{OS} and P_{ST} when agent interactions are planned by the probabilistic scheduler. Furthermore, it is much faster than P_{ST} in terms of the expected number of useful interactions (i.e., interactions that change the energy distribution in the population) needed to reach energy balance.

6 The Energy-Aware Star Network Formation Problem

In this section, we present four interaction protocols which form a star network structure in a population of (initially) disconnected agents (Fig. 2). In addition, the protocols aim at minimizing the energy distance metric which was defined in the previous sections. At any time t, two agents u, and u' are selected to interact by the probabilistic scheduler. The protocols are then executed in order to change the configuration of each agent. The protocols differ on the types of interactions that they allow, on the amount of energy that is exchanged during each interaction, and on the size of memory available on each agent. Each agent has a state from a set of states $Q = \{c, p, h_1, \ldots, h_d\}$. All agents, initially, assume that they are central agents (they are assigned the state c). The state p signifies that an agent is peripheral to the star network. The states $\{h_1, \ldots, h_d\}$ do not alter the network structure. They are used by some of the protocols in order to improve their performance.

6.1 Full Transfer P_{FT}

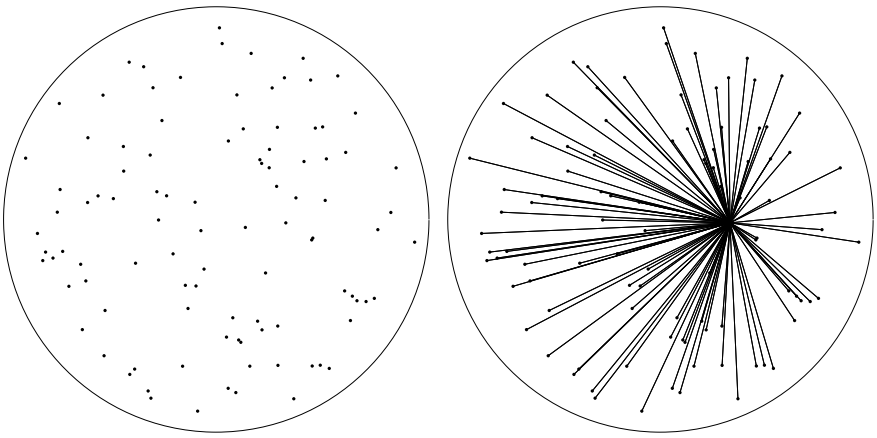
In this section, we present a straightforward protocol that can be seen as a lower bound to the performance of the other proposed protocols discussed in sections below. Protocol 4 represents the pseudo-code of P_{FT}. In this protocol, there are three main interaction cases.

1. The first case is when both agents are central. In this case, one of them will randomly be selected to remain central, and the other one will become peripheral

Protocol 4: Full Transfer P_{FT}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$ and states $q_u, q_{u'}$

```
if q_u == c AND q_u' == c then
    agent = randomly_select_agent(u, u');
    if agent == u then
        q_u' = p;
        q_{u,u'} = 1;
        ε_sent = ε_u' − E_min;
        ε_u = ε_u + ε_sent ∗ (1 − β);
        ε_u' = ε_u' − ε_sent;
    else
        q_u = p;
        q_{u,u'} = 1;
        ε_sent = ε_u − E_min;
        ε_u' = ε_u' + ε_sent ∗ (1 − β);
        ε_u = ε_u − ε_sent;
else if q_u == p AND q_u' == p then
    if q_{u,u'} == 1 then
        q_{u,u'} = 0;
else
    if q_{u,u'} == 0 then
        q_{u,u'} = 1;
```



(a) Before. (b) After.

Fig. 2 The network before and after the star construction

and it will transmit all its available energy (except from a small amount of it, denoted as E_{min}, which is needed in order for the agent to remain operational) to the central agent. In addition, a connection between them will be established.

2. In the second case, both agents are peripherals. If a connection between them exists, the algorithm removes it and no energy is transferred.
3. In the final case, one agent is considered to be central, and the other one is peripheral. In this case, if there is no connection between them, the algorithm establishes it and like in the previous case, no energy is transferred between them. That is because the agent that is peripheral has already transferred its excess energy when it was a central node and became peripheral (case 1).

We expect this protocol, to have high energy loss and energy distance from the target distribution, due to its random nature. In addition, we expect it to converge to its final energy distribution in a low number of interactions. Intuitively, this protocol will have the following outcomes:

a. *High energy distance*, because the agents transfer more energy than necessary. This means that the peripheral agents will have a lower amount of energy and the central agent will have a larger amount of it, with respect to the optimal distribution.
b. *High energy loss*. Since the selection of the agent that will become peripheral and transfer its energy to the central agent is random, it is probable to transmit energy to an agent that in a later interaction will become peripheral as well and it will have to transmit that energy again. This means that the energy transmitted by the first agent will have made many hops until it reaches the final central agent and will be affected by the energy loss factor multiple times. In addition, when two agents interact and one has higher energy level than the other, the optimal choice would be to transfer the least possible amount of energy in order to avoid the energy loss. However, this is not the case in this protocol because of the random selection mentioned above.
c. *Low number of interactions* until the star network structure (and the final energy distribution) is formed. The final energy distribution only needs $n - 1$ energy transmissions, since each agent (except the one that will remain the central agent in the final network configuration) transmits all of its energy in a single interaction. Once the central agent choice is finalized, the star structure will be finalized as well.

6.2 Half Transfer P_{HT}

In this section, we describe another interaction protocol called Half Transfer P_{HT}, which allows the agents to store their initial energy level in their memory. Each agent is only aware of its own initial energy, so this does not constitute global knowledge by the agents.

The decisions made by this protocol differ from the P_{FT} in two ways. When two central agents interact, the agent with highest energy level at that time will remain central, while the other one will become peripheral. The second difference lies in the amount of energy that is transferred when there is an energy exchange. The peripheral agent will keep half of its initial energy and will transmit the rest. This decision is based on the observation that in order to achieve the desired energy distribution, the central agent will have to acquire half of the total energy of the network. Aside from these two differences, the P_{HT} protocol operates exactly like the P_{FT} protocol.

We expect this protocol to improve on the results of the P_{FT} protocol on both the energy loss and energy distance due to the nonrandom selection of agents and the amount of energy that is transferred.

We expect this protocol to have lower energy distance than the P_{FT} protocol, since the central agent will have almost half of the total energy of the network. Moreover, the energy loss is expected to be lower since the protocol takes into account which agents have higher amount of energy to transfer.

a. *Low energy distance*. Since the central agent will have almost half of the total energy of the network, the energy distance to the desired distribution will be lower than the one from the P_{FT} protocol.
b. *High energy loss*. The choice on which agent will become peripheral is not random like in the P_{FT} protocol. Since the agents can not be sure on which agent will be the central in the final network structure, the same problem as in the P_{FT} protocol arises. More specifically, the energy will make many hops until it will arrive to the correct agent, and thus suffers of the loss factor multiple times.
c. *Low number of interactions*. Since the P_{HT} protocol works in the same way as the P_{FT} protocol, the time it needs to complete the star structure will be comparable to the time needed by the P_{FT} protocol.

6.3 Degree Aware P_{DA}

The protocol Degree Aware P_{DA} *aims to estimate the total number of agents* that are present in the network. This information is useful since each agent can adapt its energy exchanges in order to reduce the energy distance. Each central agent's estimation is the number of their connected neighbors. The noncentral agents store the maximum estimation among the agents that they have interacted with. The maximum estimation is exchanged in each interaction, thus ensuring its propagation to every agent.

In order to improve the estimation, each agent goes through d *halted states* $(h_1, h_2, \ldots, h_d)$ before it becomes peripheral; d is a parameter whose choice will be fine-tuned. The transition from a halted state to the next one is performed whenever the agent interacts with a central agent, that has a higher estimation. Please note that the next state after the final halted state h_d is the peripheral state.

To reduce energy loss, the protocol allows energy transfer only between agents where the one is central and has the highest estimation between them and the other one is peripheral (or is in the last halted state and becomes peripheral in this interaction). The selected amount of transmitted energy is described below. There are four main interaction cases in this protocol:

Protocol 5: Half Transfer P_{HT}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$, initial energies $\varepsilon_u^{initial}, \varepsilon_{u'}^{initial}$ and states $q_u, q_{u'}$

```
if q_u == c AND q_u' == c then
    agent = NULL;
    if ε_u == ε_u' then
        agent = randomly_select_agent(u, u');
    if ε_u > ε_u' OR agent == u then
        q_u' = p;
        q_{u,u'} = 1;
        ε_sent = (1/k) * (ε_u' − ε_u'^initial/2);
        ε_u = ε_u + ε_sent * (1 − β);
        ε_u' = ε_u' − ε_sent;
    else
        q_u = p;
        q_{u,u'} = 1;
        ε_sent = (1/k) * (ε_u − ε_u^initial/2);
        ε_u' = ε_u' + ε_sent * (1 − β);
        ε_u = ε_u − ε_sent;
else if q_u == p AND q_u' == p then
    if q_{u,u'} == 1 then
        q_{u,u'} = 0;
else
    if q_{u,u'} == 0 then
        q_{u,u'} = 1;
```

1. In the case where both agents are centrals, the agent with the lowest estimation becomes a first-level halted agent. A connection between them is established, the estimation of the central agent is increased by one and the estimation of the halted agent is updated to this maximum value as well.

Protocol 6: Degree Aware P_{DA}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$ and states $q_u, q_{u'}$

$r_u = number\_of\_neighbors(u)$;
$r_{u'} = number\_of\_neighbors(u')$;
$x = max\{r_u, r_{u'}\}$;
if $q_u == c\ AND\ q_{u'} == c$ **then**
  $agent = NULL$;
  if $r_u == r_{u'}$ **then**
    $agent = randomly\_select\_agent(u, u')$;
  if $r_u > r_{u'}\ OR\ agent == u$ **then**
    $q_{u'} = h_1$;
    $r_u = r_{u'} = x + 1$;
    $q_{\{u,u'\}} = 1$;
  else
    $q_u = h_1$;
    $r_u = r_{u'} = x + 1$;
    $q_{\{u,u'\}} = 1$;
else if $q_u, q_{u'} \in \{p, h_1, \ldots, h_d\}$ **then**
  $q_{\{u,u'\}} = 0$;
  $r_u = r_{u'} = x$;
else if $q_u == c\ AND\ q_{u'} \in \{p, h_1, \ldots, h_d\}$ **then**
  if $r_u \geq r_{u'}$ **then**
    if $q_{u'} \geq h_d\ OR\ q_{u'} == p$ **then**
      $q_{u'} = p$;
      if $q_{\{u,u'\}} == 0$ **then**
        $q_{\{u,u'\}} = 1$;
        $r_u = r_u + 1$;
      $\varepsilon_{sent} = \frac{1}{k} * \frac{\varepsilon_u}{\varepsilon_u + \varepsilon_{u'}} * \varepsilon_{u'}$;
      if $\varepsilon_u < \varepsilon_{u'} * r_u$ **then**
        $\varepsilon_u = \varepsilon_u + \varepsilon_{sent} * (1 - \beta)$;
        $\varepsilon_{u'} = \varepsilon_{u'} - \varepsilon_{sent}$;
      else if $\varepsilon_u > \varepsilon_{u'} * r_u$ **then**
        $\varepsilon_{u'} = \varepsilon_{u'} + \varepsilon_{sent} * (1 - \beta)$;
        $\varepsilon_u = \varepsilon_u - \varepsilon_{sent}$;
      end if
    else
      $q_{u'} = h_{i+1}$;
    $r_{u'} = r_u$;
  else
    $q_u = h_1$;
    $r_u = r_{u'}$;
else if $q_u \in \{p, h_1, \ldots, h_d\}\ AND\ q_{u'} == c$ **then**
  Similarly with the case above by symmetry.
end if

2. In the case where each agent is either peripheral or halted, the agents exchange the maximum estimation and delete their connection if it exists.
3. In the case where one agent is central and the other one is peripheral, if the central agent has lower estimation than the peripheral, it becomes a first-level halted agent and updates its estimation. Otherwise, their states remain the same, but if the agents are not connected, they establish a connection and update their estimations to the new maximum one. If the energy level of an agent times the maximum estimation (between these two agents) is larger than the energy level of the other agent, it will transmit an amount of energy that is equal to $(1/k) \times (E(u) \times E(u'))/E(u) + E(u'))$ where parameter k *is used to limit the energy exchanged between the agents and thus the energy loss*, when the estimation is not equal to the actual network size.
4. In the fourth case where one agent is central and the other one is halted, if the halted agent has larger estimation than the central, the latter becomes halted as well. Else, if the central agent has the larger estimation but the level of the halted agents is not d, i.e., its state is not the h_d, it moves to the next level halted state. Otherwise, if the central agent has the highest estimation and the halted agent's state is the last one, then the halted agent becomes peripheral. If the agents are not connected, they establish a connection and increase the estimation of the central agent by one. If the energy level of an agent times the maximum estimation (between these two agents) is larger than the energy level of the other agent, it will transmit to it an amount of energy that is equal to $(1/k) \times (E(u) \times E(u'))/(E(u) + E(u'))$. Also, both agents will update their estimation to the maximum one between them.

Intuitively, this protocol is meant to have the following performance:

a. *Almost zero energy distance* is expected by this protocol since the agents estimate the size of the network and thus they adapt their energy exchanges.
b. *Low energy loss* is assumed to be achieved as the protocol makes more targeted energy exchanges, with the usage of the halted states discussed above. Thus, most of the energy is transferred directly to the final central agent of the network avoiding energy loss through multiple hops. *The performance depends on the size of* d. *The larger this parameter, the higher the number of additional states is and thus, a lower energy loss is expected.*
c. *High number of interactions* until the global star is completed since the agents need to go through d additional states until they become peripherals. The higher the parameter d, the higher the number of required interactions is.

In other words, the protocol achieves *a tunable trade-off* between energy efficiency and convergence time.

6.4 Fully Adaptive P_{FA}

The protocol Fully Adaptive P_{FA} aims to improve on the ideas of P_{DA} protocol introduced in the previous section. The protocol assumes slightly stronger agents with the ability to store more information on their memory. In addition to storing the estimation of the network size, as discussed above, each agent also stores the energy level (at the time of their interaction) of the last central agent (e_c) it has interacted with.

This protocol works in the same way as the P_{DA} protocol. The main difference lies in the way the agents exchange energy. There are two different types of energy exchanges.

a. When a central agent (u) interacts with either a peripheral or a d-level halted agent (u'), the energy to be exchanged between them is calculated with this formula $\varepsilon_{sent} = (1/k) \times (E(u)(r(u) - 1) - E(u'))/(r(u) + 1)$. When this value is negative (respectively positive), it means that u has less (respectively more) energy than it is required in order to achieve the desired energy distribution and thus, it receives (respectively transmits) that (absolute) amount of energy from (respectively to) u'.
b. When two peripheral agents (u, u') interact, before they exchange energy, they attempt to find the optimal energy level a peripheral agent should have according to the desired energy distribution. This is done using the stored value for the energy of the last central agent they have interacted with and is defined as $e_p(u/u') = e^c/(r(u/u') - 1)$. Both agents calculate this value and they exchange energy if and only if they are on opposite sides of both these calculated values (i.e., $E(u) > e_p(u)$ & $E(u') < e_p(u)$ & $E(u) > e_p(v)$ & $E(u') < e_p(u')$). If all these conditions are true, then the agent with the highest energy level (e.g., agent u) transmits energy according to the following formula: $\varepsilon_{sent} = (1/k) \times (E(u') - E(u))/2$. As in the previous protocol, k is used to limit the energy exchanged between the agents and thus the energy loss, when the estimation is not equal to the actual network size.

This protocol aims to improve the outcomes of the P_{DA} protocol, especially in the energy loss metric because of the way that the energy is exchanged between agents.

Protocol 7: Fully Adaptive P_{FA}

Input : Agents u, u' with energy levels $\varepsilon_u, \varepsilon_{u'}$ and states $q_u, q_{u'}$

$r_u = number\_of\_neighbors(u)$;
$r_{u'} = number\_of\_neighbors(u')$;
$x = max\{r_u, r_{u'}\}$;
if $q_u == c\ AND\ q_{u'} == c$ **then**
  $agent = NULL$;
  if $r_u == r_{u'}$ **then**
    $agent = randomly\_select\_agent(u, u')$;
  if $r_u > r_{u'}\ OR\ agent == u$ **then**
    $q_{u'} = h_1; r_u = r_{u'} = x + 1; q_{\{u,u'\}} = 1; e^c_u = \varepsilon_u$;
  else
    $q_u = h_1; r_u = r_{u'} = x + 1; q_{\{u,u'\}} = 1; e^c_u = \varepsilon_{u'}$;
else if $q_u == p\ AND\ q_{u'} == p$ **then**
  $q_{\{u,u'\}} = 0; r_u = r_{u'} = x$;
  $m_u = e^c_u / r_u; m_{u'} = e^c_u / r_{u'}$;
  $\varepsilon_{sent} = \frac{1}{k} * |\frac{\varepsilon_u - \varepsilon_{u'}}{2}|$;
  if $\varepsilon_u > e^u_p\ AND\ \varepsilon_{u'} < e^u_p$ **then**
    if $\varepsilon_u > e^v_p\ AND\ \varepsilon_{u'} < e^v_p$ **then**
      $\varepsilon_{u'} = \varepsilon_{u'} + \varepsilon_{sent} * (1 - \beta); \varepsilon_u = \varepsilon_u - \varepsilon_{sent}$;
  else if $\varepsilon_{u'} > e^u_p\ AND\ \varepsilon_u < e^u_p$ **then**
    if $\varepsilon_{u'} > e^v_p\ AND\ \varepsilon_u < e^v_p$ **then**
      $\varepsilon_u = \varepsilon_u + \varepsilon_{sent} * (1 - \beta); \varepsilon_{u'} = \varepsilon_{u'} - \varepsilon_{sent}$;
  end if
else if $q_{u/u'} \in \{p, h_1, \ldots, h_d\}\ AND\ q_{u'/u} \in \{h_1, \ldots, h_d\}$ **then**
  $q_{\{u,u'\}} = 0; r_u = r_{u'} = x$;
else if $q_u == c\ AND\ q_{u'} \in \{p, h_1, \ldots, h_d\}$ **then**
  if $r_u \geq r_{u'}$ **then**
    if $q_{u'} \geq h_d\ OR\ q_{u'} == p$ **then**
      $q_{u'} = p; e^c_{u'} = \varepsilon_u$;
      if $q_{\{u,u'\}} == 0$ **then**
        $q_{\{u,u'\}} = 1; r_u = r_u + 1$;
      $e^c_{u'} = \varepsilon_u; \varepsilon_{sent} = \frac{1}{k} * |\frac{\varepsilon_{u'} * r_u - \varepsilon_u}{r_u + 1}|$;
      if $\varepsilon_{sent} < 0$ **then**
        $\varepsilon_{u'} = \varepsilon_{u'} + \varepsilon_{sent} * (1 - \beta)$;
        $\varepsilon_u = \varepsilon_u - \varepsilon_{sent}$;
      else if $\varepsilon_{sent} > 0$ **then**
        $\varepsilon_u = \varepsilon_u + \varepsilon_{sent} * (1 - \beta)$;
        $\varepsilon_{u'} = \varepsilon_{u'} - \varepsilon_{sent}$;
      end if
    else
      $q_{u'} = h_{i+1}$;
    $r_{u'} = r_u$;
  else
    $q_u = h_1; r_u = r_{u'}$;
else if $q_u \in \{p, h_1, \ldots, h_d\}\ AND\ q_{u'} == c$ **then**
  Similarly with the above case by symmetry.
end if

7 Performance Evaluation

We conducted simulations in order to evaluate the performance of the proposed protocols. The simulation environment is Matlab R2016a.

For statistical smoothness, we conducted each simulation 100 times. The statistical analysis of the findings (the median, the lower and upper quartiles, snd outliers of the samples) demonstrates very high concentration around the mean and so, in the following simulation results, we depict only the average values.

7.1 The Population Energy Balance Problem

In this section, we present the evaluation of the proposed protocols for the population energy balance problem. More specifically, we compared the protocols P_{OS}, P_{ST} and P_{OA} by conducting experiment runs of 1.000 useful interactions, where the nodes to interact are selected by a probabilistic scheduler. We assign an initial energy level value to every agent of a population consisting of $|m| = 100$ agents uniformly at random, with maximum battery cell capacity 100 units of energy. The constant β of the loss function is set to three different values, as different energy losses might lead to different performance (see Fig. 3).

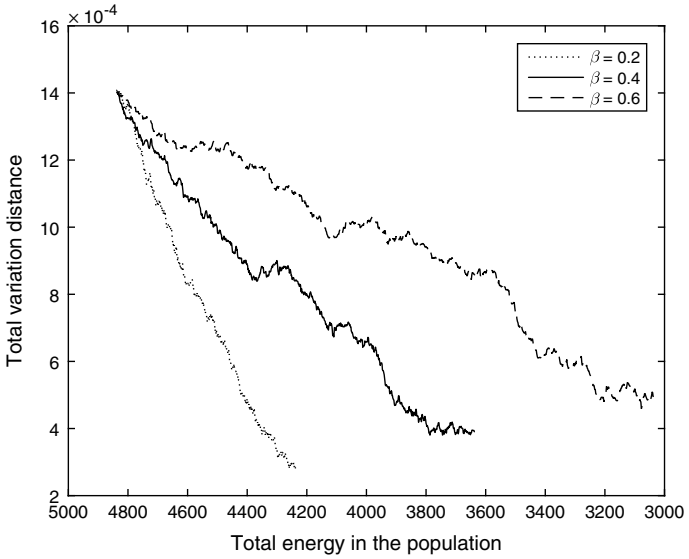


Fig. 3 Performance of P_{ST} for different values of β. Different loss functions affect the performance of the protocol

7.1.1 The Impact of β

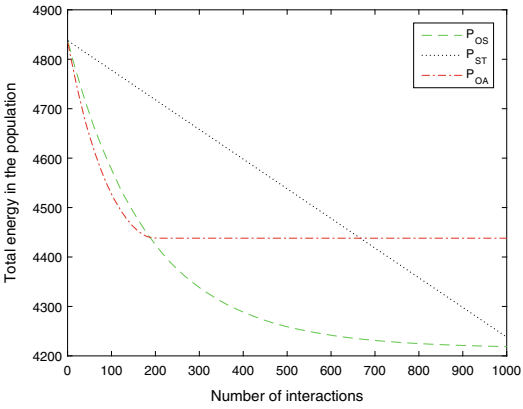
Different loss functions $L(\varepsilon)$ lead to different performance of the interaction protocols, both when running the same protocol and when comparing different protocols. Regarding the impact of different values of the β constant on the same protocol, an example is shown in Fig. 3. The total variation distance w.r.t. the remaining energy in the population is shown. We ran the P_{ST} protocol for values 0.2, 0.4, and 0.6. The results clearly show that the bigger the β, the larger the variation distance for a given total level of energy in the population. For this reason, we decided to comparatively evaluate our protocols for different values of β, as shown in Figs. 4, 5, and 6. As for the impact on different protocols, if we observe Fig. 4 carefully, we can see that, for the same total initial energy and number of useful interactions, when the β constant and consequently the energy loss increases, the rate of total energy loss also does.

7.1.2 Energy Loss

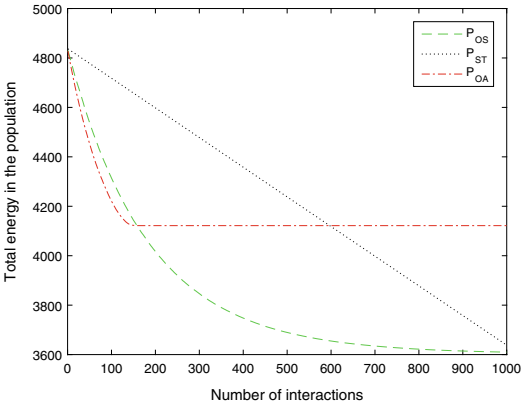
Figure 4 are depicting the total energy of the population over time, for 1.000 useful agent interactions of the three protocols P_{OS}, P_{ST} and P_{OA}. Each protocol's behavior is similar, regardless of the value of β, but with higher losses when β increases. The energy loss rate for P_{OS} and P_{OA} is high in the beginning, until a point of time when energy stops leaking outside the population. This is explained by the fact that both protocols perform interactions of high energy transfer amounts ε which lead to high $L(\varepsilon)$. After those interactions, P_{OS} performs energy transfers of very small ε forcing the energy loss rate to drop sharply and P_{OA} drives the energy levels of most agents to the same side of the average value, rendering useful interactions very rare. P_{ST} has a smoother, linear energy loss rate since ε is a very small fixed value.

7.1.3 Energy Balance

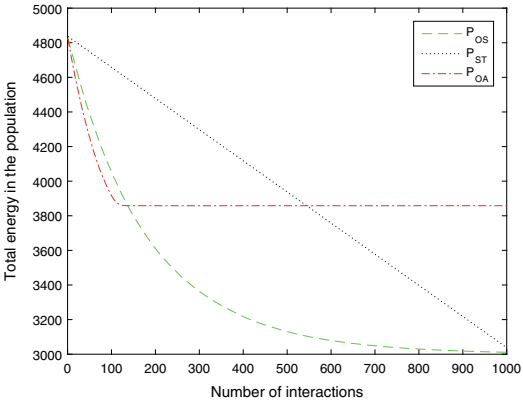
Useful conclusions about energy balance of the population can be derived from the total variation distance over time depiction in Fig. 5. A first remark is that the protocols are balancing the available energy in the population in an analogous rate to the energy loss rate. Since a better energy balance is expressed by lower values of total variation distance, it is apparent that eventually the best balance after 1.000 useful interactions is provided by P_{OS}. However, note that this is a conclusion regarding only the energy balance, not taking into account the losses from the charging procedure. As we will see in the next subsection, better balance does not necessarily lead to higher overall efficiency, w.r.t. energy loss. If we take a better look at the energy balance figures, we observe that even if the total variation distance follows a decreasing pattern, it is not strictly decreasing. This is natural since many interactions can temporarily lead to a worse energy balance in the population due to sharp changes in the distribution of total energy.



(a) Energyloss, $\beta = 0.2$.

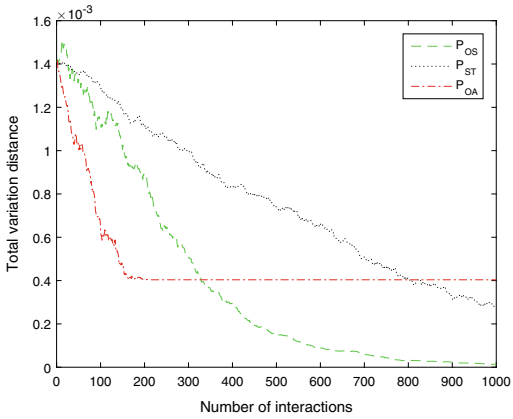


(b) Energyloss, $\beta = 0.4$.

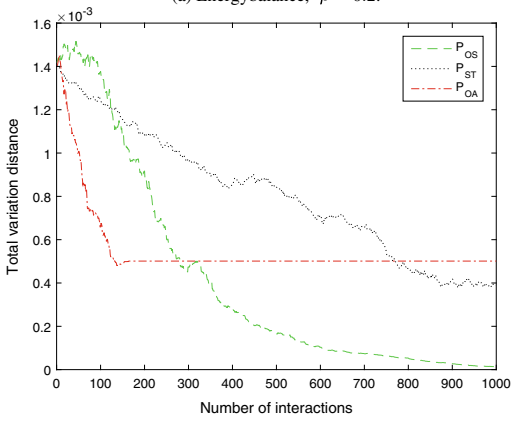


(c) Energyloss, $\beta = 0.6$.

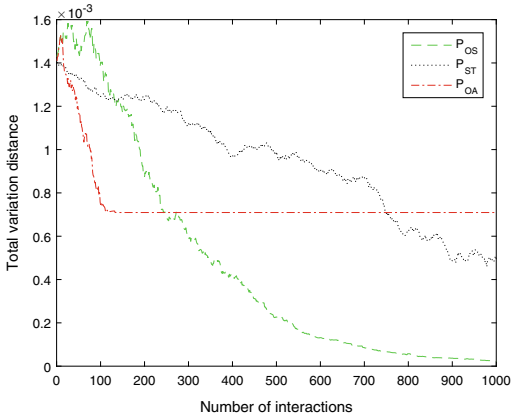
Fig. 4 Comparison of the three protocols on energy loss metric for different values of β



(a) Energybalance, $\beta = 0.2$.



(b) Energybalance, $\beta = 0.4$.



(c) Energybalance, $\beta = 0.6$.

Fig. 5 Comparison of the three protocols on energy balance metric for different values of β

7.1.4 Running Time

The time that each protocol needs for balancing the available energy in the population, is not a negligible factor. Quick balancing leads to transfers of significantly smaller amounts of energy among agents and consequently to lower energy losses. On the other hand, in order to achieve quick balancing, in some cases, there has been already much energy loss due to frequent lossy interactions. In Figs. 4, 5, and 6, we can see that P_{OA} is the fastest to achieve a stable level of energy balance in the population, as opposed to P_{OS}, which is wasteful in terms of running time. P_{ST} timing performance lies somewhere in between the two other protocols, since it is able to conduct all types of interactions (unlike P_{OA} in which only some interactions are allowed and P_{ST} which performs only interactions of small ε).

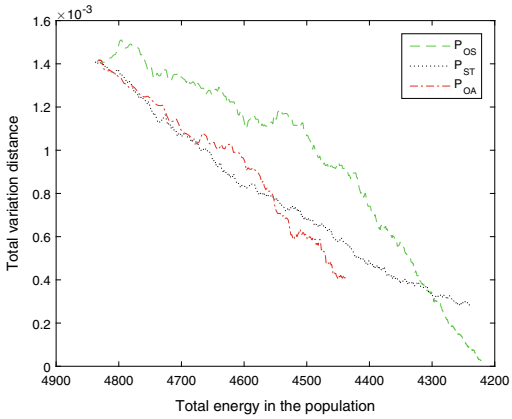
7.1.5 Overall Efficiency

We measure the overall efficiency of a protocol by taking into account the both energy losses and energy balance in the population. This combination of the two crucial properties is shown in Fig. 6a–c, where P_{ST} and P_{OA} clearly outperform P_{OS}, most of the time. More specifically, although P_{OS} achieves very good balance quickly, the impact of energy loss affects very negatively its performance. This pattern results in the fact that for the same amount of total energy in the population, P_{ST} and P_{OA} achieve better total variation distance than P_{OS}. It is also clear that eventually, P_{OA} outperforms both P_{OS} and P_{ST} when agent interactions are planned by the probabilistic scheduler. Furthermore, it is much faster than P_{ST} in terms of the number of useful interactions.

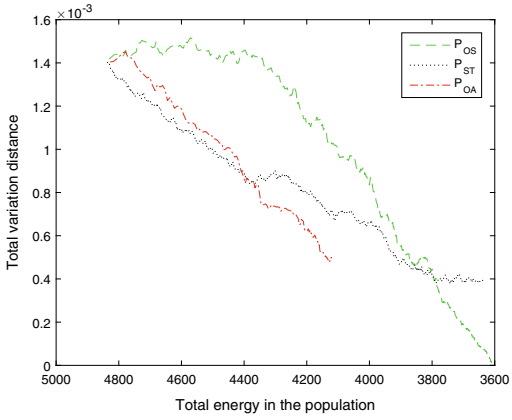
7.2 The Energy-Aware Star Network Formation Problem

In this section, we present the evaluation of the proposed protocols for the energy-aware star network formation problem. In order to abstract the real network of diverse portable devices, we apply a nonuniform initial energy distribution among the agents. The number of agents varies from 20 to 100. The total initial energy is analogous to the number of agents. More specifically, the total energy is set to $3000 \cdot [20 : 20 : 100]$ for 20, 40, 60, 80, and 100 agents, respectively. Our protocols are designed to run constantly, but for the purposes of this chapter, we plot their performance until the desired energy distance is achieved. The wireless energy loss factor β is different at each interaction and follows the Normal Distribution. More specifically, $\beta \sim N(0.2, 0.05)$.

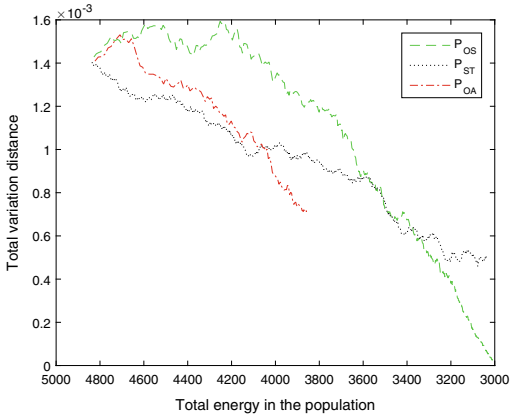
In this section, we provide our simulation results on various metrics. At first, we find the best value of the parameters d and k for various metrics and various number of agents. In order to select the best values for d and k for each protocol, we design *a metric that takes into account both the energy loss and the speed of each protocol*



(a) Efficiency, $\beta = 0.2$.



(b) Efficiency, $\beta = 0.4$.



(c) Efficiency, $\beta = 0.6$.

Fig. 6 Comparison of the three protocols on efficiency for different values of β

trying to optimize their trade-off. This metric is defined as follows:

$$y = t \times energy\_loss_t \times energy\_distance_t \tag{35}$$

where t is the time when the protocol with the worst performance reaches its best energy distance and the factors $energy\_distance_t$ and $energy\_loss_t$ are the energy distance and total energy loss of each protocol at time t, respectively. The values of d and k which give the minimum value for y are selected.

These selected values are used to investigate the protocols' performance on the following metrics: (a) structural distance, (b) energy distance, (c) energy loss, and (d) speed. The metrics (a) and (b) are already described in Sect. 4. Metric (c) refers to the amount of energy lost due to energy exchanges and metric (d) represents how fast, i.e., the number of interactions, the protocols achieve the desired energy distance.

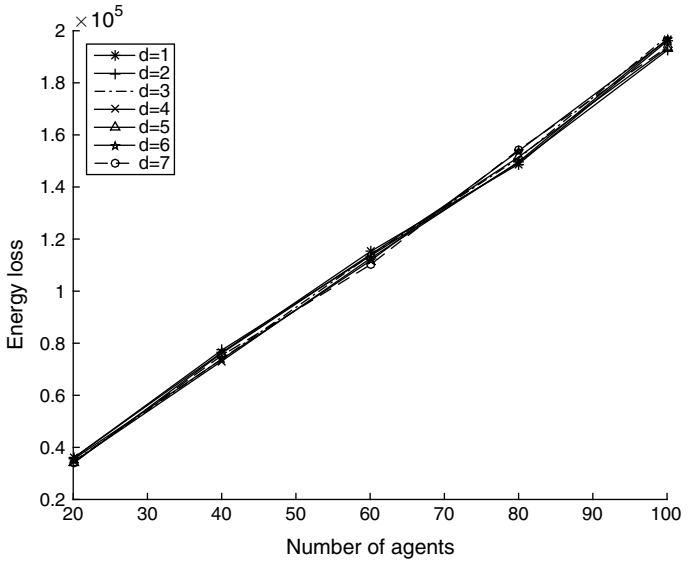
7.2.1 Fine Tuning of Parameter D

In this section, we conduct simulations in order to evaluate the performance of the P_{DA} and the P_{FA} protocols for different values of the parameter d which indicates the number of halted states an agent will have to pass through before its state becomes peripheral. We select various number of agents (20, 40, 60, 80, and 100) and we set the value of k, which is used to withhold the amount of transmitted energy, to $k = 1$ in order to evaluate the effect of d independently of k.

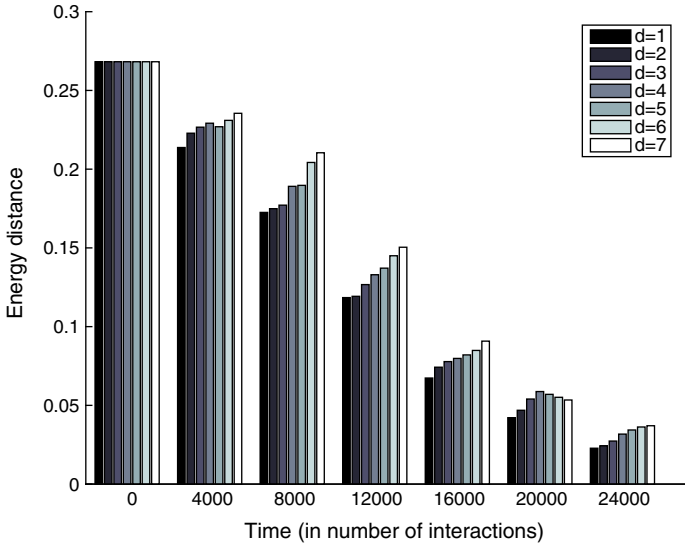
Degree-Aware protocol. Figure 7 presents the effect of d on the P_{DA} protocol. More specifically, Fig. 7a depicts the impact of d on the total energy loss when the protocol achieved an energy distance equal to 0.2. Interestingly, the energy loss is not affected by the value of d. This means that even a value of $d = 1$ is sufficient for the P_{DA} protocol to reach its optimal energy loss.

In order to further evaluate the effect of d, we also need to take into account its effect on the speed of the protocol (i.e., how many interactions are required to achieve the desired energy distribution). In Fig. 7b, as expected, we observe that the lower the value of d the faster the P_{DA} protocol achieves the energy distribution. This is natural since the agent does not have to pass through many halted states in order to begin transmitting its energy to prospective central agents.

Fully Adaptive protocol. Figure 8 presents the effect of d on the P_{FA} protocol. More specifically, Fig. 8a depicts the impact of d on the total energy loss. We observe that for lower values of d the total energy loss is higher. This is explained by the fact that an agent, when in a halted state, will not make any energy exchanges. The more halted states the agent has to pass through, the higher the confidence will be about the estimated size of the network, thus making any energy exchanges more precise. We also observe that for larger network size, the total energy loss is higher. This is expected because the number of energy exchanges is much larger. Another observation that can be made by Fig. 8a is that for values of $d \geq 3$ the effect on the energy loss is diminished.

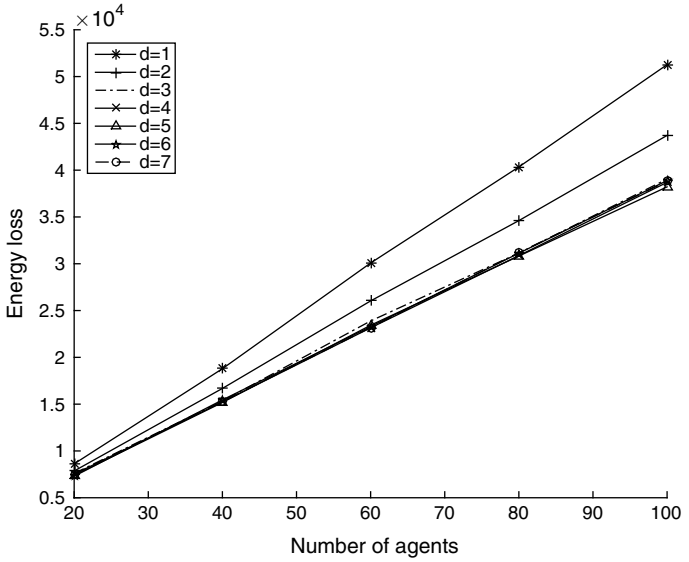


(a) Energy loss to total number of agents.

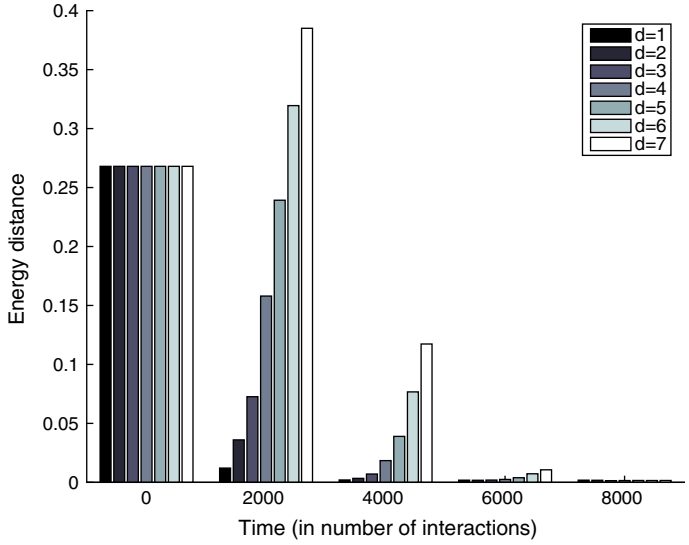


(b) Energy distance to number of interactions.

Fig. 7 Evaluation of the parameter d for the P_{DA} protocol



(a) Energy loss to total number of agents.



(b) Energy distance to number of interactions.

Fig. 8 Evaluation of the parameter d for the P_{FA} protocol

We further investigate the effect of d on the speed of the protocol. As in the P_{DA} protocol, in Fig. 8b, we observe that the lower the value of d, the faster the P_{FA} protocol achieves the targeted energy distribution. In fact, we see that for $d \geq 6$ the P_{FA} protocol, in the first interactions, increases the energy distance significantly. This can be explained by the fact that the structural distance is being decreased, but the necessary energy exchanges are not being performed. We will refer to the period, in the lifetime of a protocol, during which there are many fluctuations in the energy distance metric as *metastability period*.

After the execution of the protocols, we calculate the value d from the Eq. 35. The values for both the P_{DA} and the P_{FA} protocol are $d(P_{FA}) = d(P_{DA}) = 1$.

7.2.2 Fine Tuning of Parameter K

In this section, we conduct simulations in order to evaluate the performance of the P_{HT}, the P_{DA}, and the P_{FA} protocols for different values of the parameter k which is used to withhold the amount of transmitted energy. More specifically, if in any given interaction, an agent is supposed to transmit amount of energy e_x, with the addition of k it will transmit $\frac{e_x}{k}$. In order to evaluate the effect of k solely, we set the value of the parameter $d = 1$. As in the previous section, we conduct simulations with a various number of network sizes, comprised of 20, 40, 60, 80, and 100 agents, respectively.

Similar to the parameter d, we evaluate the effect of k on the total energy loss as well as the speed in which the protocols achieve an energy distance close to the desired distribution.

Half Transfer protocol. Figure 9a presents the effect of k on the energy loss factor for all network sizes. We observe that the effect of k increases as the network size increases. In addition, with higher values of k the P_{HT} protocol achieves lower energy loss. However, as shown in Fig. 9b, the protocol arrives faster to a lower energy distance with lower values of k.

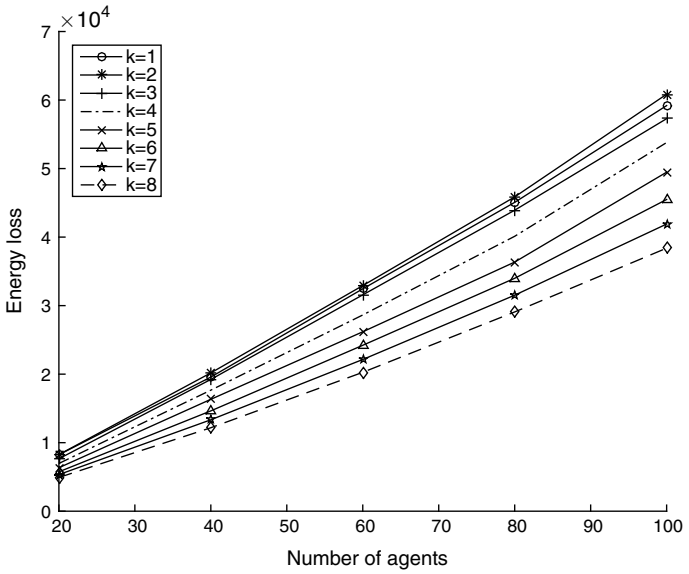
Degree-Aware protocol. Figure 10a presents the effect of k on the energy loss factor. We observe that with higher values of k the P_{DA} protocol, achieves lower energy loss. This is expected because during the first few energy exchanges, we are in the metastability period described above. The performed energy exchanges during this period are not optimal.

Figure 10b depicts the effect of k on the speed of the P_{DA} protocol. The results clearly show that k's effect is minuscule and can be dismissed.

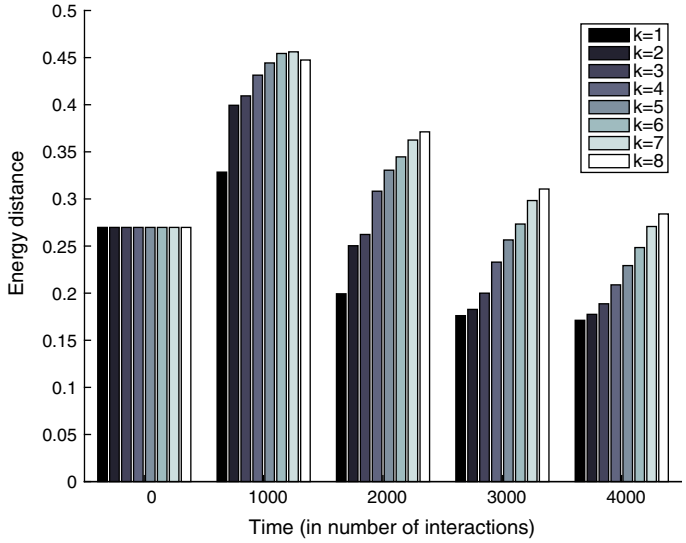
Fully Adaptive protocol. In Fig. 11a, similarly to the P_{DA} and P_{HT} protocols, it is observed that higher values of k lead to lower energy loss. It is worth noting that for values of $k \geq 4$, the performance of the protocol is similar with respect to this metric.

In Fig. 11b, we can clearly see that for lower values of k the P_{FA} protocol reaches energy distance close to the desired energy distribution much faster.

Similarly to the parameter d, using Eq. 35, we find the optimal value of k for each protocol. More specifically, we select $k(P_{HT}) = 1, k(P_{DA}) = 7, k(P_{FA}) = 1$.

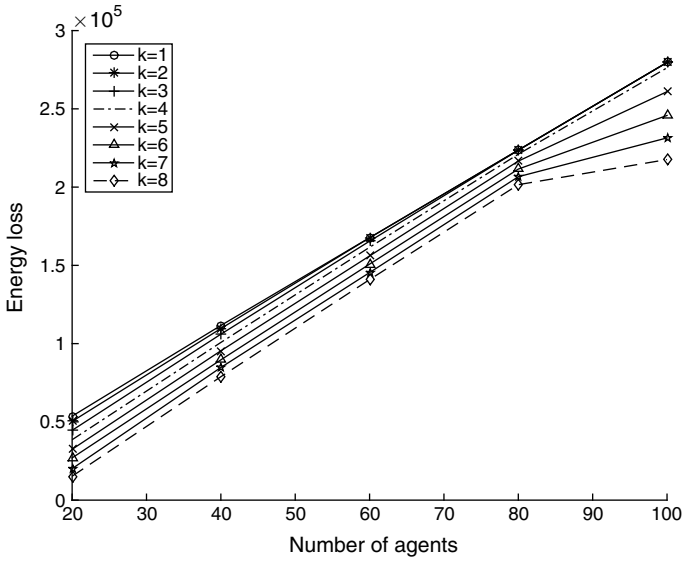


(a) Energy loss to total number of agents.

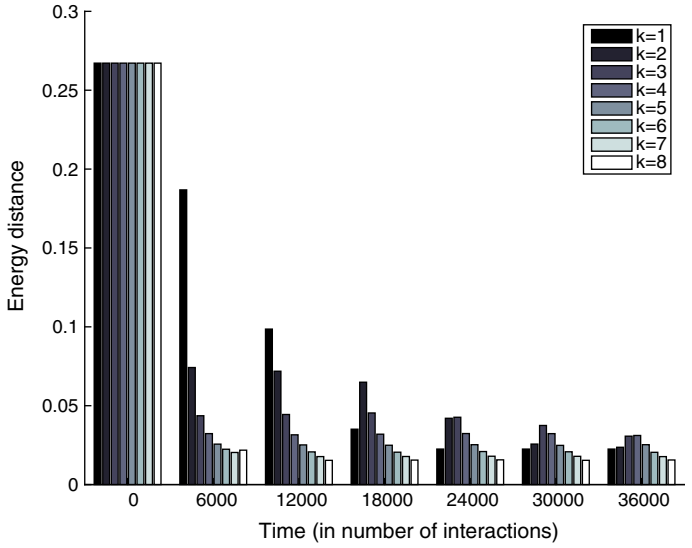


(b) Energy distance to number of interactions.

Fig. 9 Evaluation of the parameter k for the P_{HT} protocol

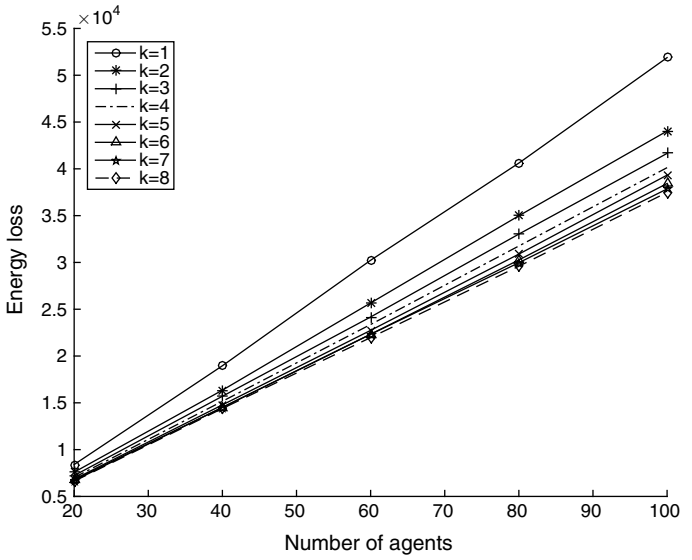


(a) Energy loss to total number of agents.

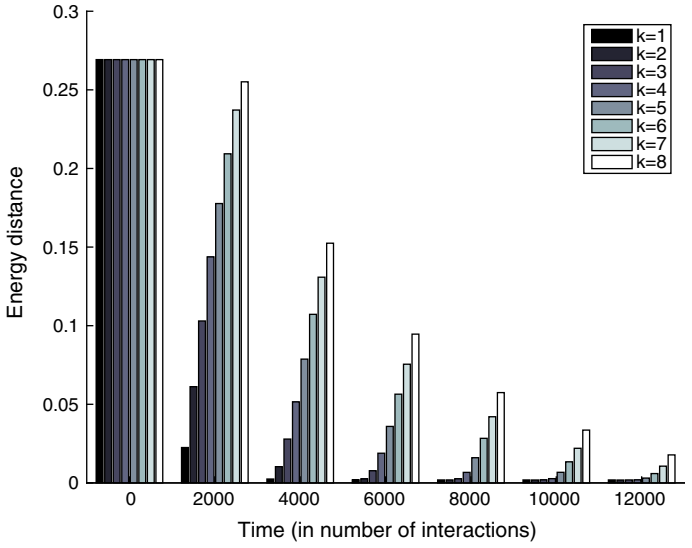


(b) Energy distance to number of interactions.

Fig. 10 Evaluation of the parameter k for the P_{DA} protocol



(a) Energy loss to total number of agents.



(b) Energy distance to number of interactions.

Fig. 11 Evaluation of the parameter k for the P_{FA} protocol

In the following sections, we present the performance of the four protocols described in the previous sections, after the fine tuning of the various parameters. We conducted simulations with different network sizes, i.e., with 20, 60, and 100 agents, respectively. We observed that each protocol has similar performance for each network size. Thus, we select to present the results for a network with 100 agents.

7.2.3 Protocols' Performance on Time to Converge

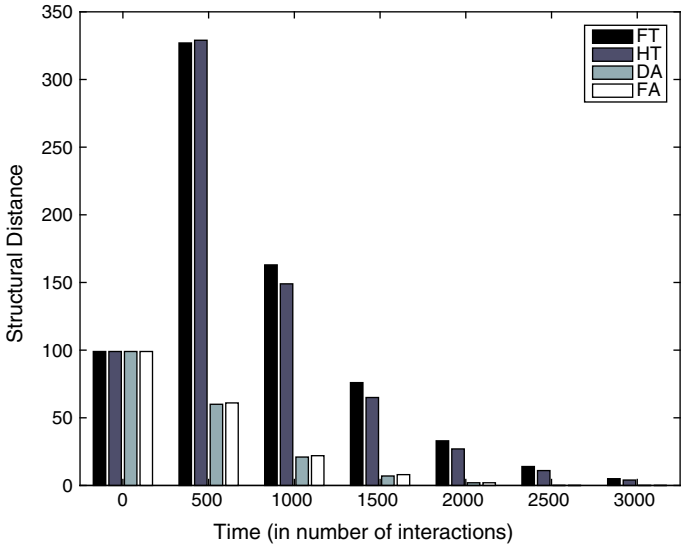
In this section, we compare the protocols performance on the number of interactions they need to build a global star network structure, as well as to achieve a low energy distance. In Fig. 12a, we can clearly see that during the metastability period the P_{FT} and P_{HT} protocols increase the structural distance. This is explained by the fact that the agents do not have any knowledge on the network size. Whenever a central agent interacts with a peripheral agent, a connection will be established, resulting in a large number of unnecessary connections between agents. When the metastability period ends, these protocols will eventually build a star network as well. The P_{DA} and P_{FA} protocols have similar performance in this metric as they build the structure relatively quickly, compared to the other two protocols. The performance gap between the two types of protocols can be explained by the power of two choices [36] that the halted states provide.

Figure 12b depicts the performance of the protocols on the number of interactions needed in order to achieve a relatively low energy distance. We clearly observe that the P_{FA} protocol, outperforms all other protocols. It reaches almost zero energy distance in relatively few interactions. The P_{DA} protocol also reaches the desired energy distribution but does so with almost double number of interactions than the P_{FA}. The P_{FT} and P_{HT} protocols do not achieve a good energy distribution. They achieve their best energy distance with relatively few interactions but that energy distance is far from the desired distribution.

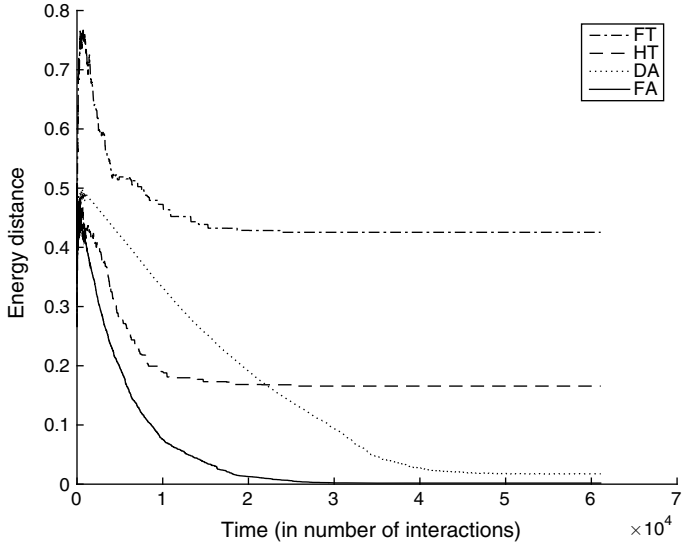
7.2.4 Protocols' Overall Performance

In this section, we perform simulations in order to compare the performance of the protocols with respect to the energy they spend in order to achieve the target energy distribution as well as their performance on how close they come to that distribution.

Figure 13a depicts the total energy lost during the energy exchanges by each protocol with respect to the total initial amount of available energy in the network. As expected, the P_{FA} protocol achieves the lowest energy loss since the energy exchanges made are more precise and focused. In order to make the comparison fair, the amount of lost energy depicted for the P_{DA} protocol is the value when the protocol reaches sufficiently low energy distance (0.05). The P_{HT} protocol also has a similar performance in this metric. In contrast, the P_{FT} protocol has the worst

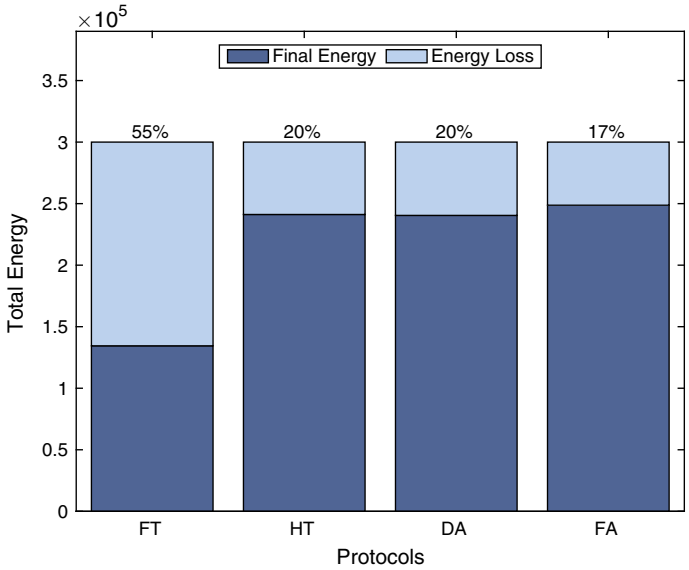


(a) Structural Distance to number of interactions.

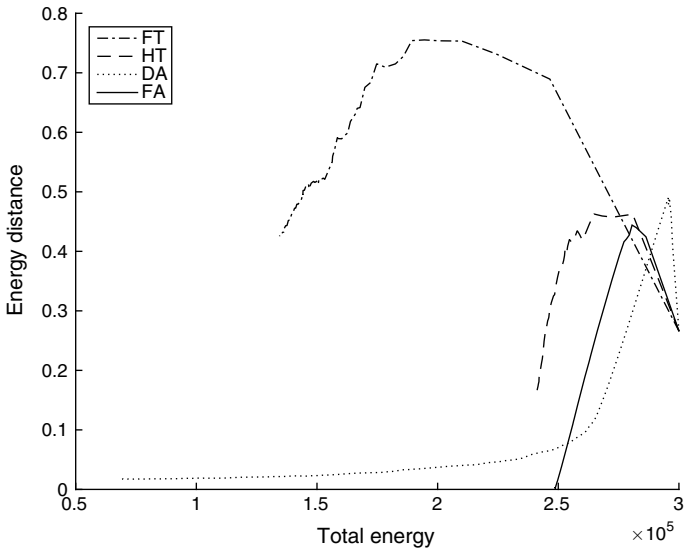


(b) Energy distance to number of interactions.

Fig. 12 Speed comparison of the different protocols



(a) Energy loss to total energy.



(b) Energy distance to total energy.

Fig. 13 Energy comparison of the different protocols

performance, spending more than half of the initial energy. This is expected due to the completely random nature of the protocol.

In Fig. 13b, we observe that during the metastability period, the P_{DA} protocol is actually better than the P_{FA}. This can be explained by the value of $k = 7$ that was selected. After this period, the P_{FA} clearly outperforms all protocols on both the energy distance metric as well as the energy loss. The P_{DA} protocol approaches the desired energy distribution but in doing so, it spends all the available energy in the network. The P_{HT} protocol manages to reduce the energy distance but it fails in approaching the P_{DA} protocols. As expected, the P_{FT} protocol performs badly in this metric as well.

7.2.5 The Loss-Less Case

Finally, we conduct simulations in order to evaluate the performance of the protocols in the loss-less case. The only metric that is affected by the loss factor is the energy distance. In Fig. 14, we observe that in the loss-less case as well, the P_{FA} protocol clearly outperforms all the other protocols. The P_{DA} protocol even though it approaches the desired energy distribution, it does so with more interactions. The P_{HT} protocol reaches a very good energy distance (0.1) in less interactions than the P_{DA} protocol but it cannot further approach the target distribution. The P_{FT} protocol is outperformed by all other protocols.

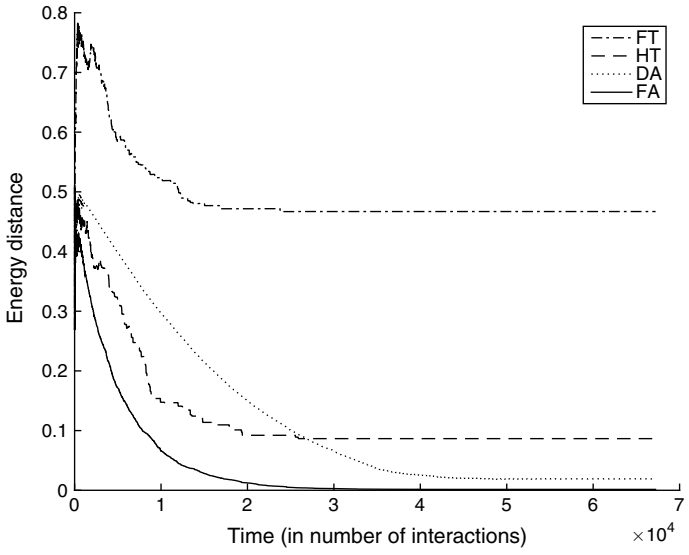


Fig. 14 Energy distance to number of interactions

Furthermore, we observe that the P_{HT} protocol, performs significantly better in the loss-less case. This is expected since the peripheral agents only keep half of their initial energy and transmit the rest to the central agent. This leads to the central agent having (almost) half of the network's energy. Due to the nonuniform distribution of the initial energies, the peripheral agents will not have the desired energy distribution. An energy balancing scheme between the peripherals was implemented in order to further reduce the energy distance, but after simulations, we observed that this only works in the loss-less case. In the case where there is energy loss, the distance actually increases since the peripheral agents will lose even more energy while trying to achieve energy balance.

8 Conclusion

In this chapter, we studied two main problems on the new topic of interactive wireless charging in populations of resource-limited, mobile agents, namely the energy balance and the energy-aware network formation (particularly for the creating a star structure). We considered both the lossless and lossy cases of energy transfer. Three protocols for the problem of energy balance between the network agents and an upper bound on the time needed to reach energy balance are provided. In addition, four interaction protocols have been proposed for the problem of energy-aware star network formation. These protocols assume different amounts of knowledge of the network and achieve different trade-offs between energy balance, time and energy efficiency.

We plan to further fine-tune the assumptions of this chapter, by considering unique features of wireless networks, e.g., wireless channel models, data communication specifics, PHY and link layer, for verifying the applicability and the practicality of the proposed concepts.

Acknowledgements This research has been cofinanced by the European Union (European Social Fund—ESF) and Greek national funds through the action entitled "Strengthening Human Resources Research Potential via Doctorate Research" of State Scholarships Foundation (IKY), in the framework of the Operational Programme "Human Resources Development Program, Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) 2014 2020.

References

1. Aldous, D., Fill, J.A.: Reversible markov chains and random walks on graphs (2002). Unfinished monograph, recompiled 2014. http://www.stat.berkeley.edu/~aldous/RWG/book.html
2. Angelopoulos, C.M., Buwaya, J., Evangelatos, O., Rolim, J.: Traversal strategies for wireless power transfer in mobile ad-hoc networks. In: Proceedings of the 18th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems, MSWiM '15, pp. 31–40. ACM, New York, NY, USA (2015). https://doi.org/10.1145/2811587.2811603

3. Angluin, D., Aspnes, J., Diamadi, Z., Fischer, M.J., Peralta, R.: Computation in networks of passively mobile finite-state sensors. In: Proceedings of the Twenty-Third Annual ACM Symposium on Principles of Distributed Computing, PODC '04, pp. 290–299. ACM, New York, NY, USA (2004)
4. Angluin, D., Aspnes, J., Eisenstat, D.: Stably computable predicates are semilinear. In: Proceedings of the Twenty-Fifth Annual ACM Symposium on Principles of Distributed Computing, PODC '06, pp. 292–299. ACM, New York, NY, USA (2006)
5. Angluin, D., Aspnes, J., Eisenstat, D.: Fast computation by population protocols with a leader. Distrib. Comput. **21**(3), 183–199 (2008)
6. Aspnes, J., Ruppert, E.: An Introduction to Population Protocols, pp. 97–120. Springer, Berlin, Heidelberg (2009)
7. Costanzo, A., Dionigi, M., Masotti, D., Mongiardo, M., Monti, G., Tarricone, L., Sorrentino, R.: Electromagnetic energy harvesting and wireless power transmission: a unified approach. Proc. IEEE **102**(11), 1692–1711 (2014)
8. Dai, H., Wu, X., Chen, G., Xu, L., Lin, S.: Minimizing the number of mobile chargers for large-scale wireless rechargeable sensor networks. Comput. Commun. (2014)
9. Del Prete, M., Berra, F., Costanzo, A., Masotti, D.: Exploitation of a dual-band cell phone antenna for near-field WPT. In: 2015 IEEE Wireless Power Transfer Conference (WPTC) (2015)
10. Griffin, B., Detweiler, C.: Resonant wireless power transfer to ground sensors from a UAV. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 2660–2665 (2012)
11. Gunduz, D., Stamatiou, K., Michelusi, N., Zorzi, M.: Designing intelligent energy harvesting communication systems. IEEE Commun. Mag. **52**(1), 210–216 (2014)
12. Guo, S., Wang, C., Yang, Y.: Mobile data gathering with wireless energy replenishment in rechargeable sensor networks. In: INFOCOM, 2013 Proceedings IEEE, pp. 1932–1940 (2013)
13. Guo, S., Wang, C., Yang, Y.: Joint mobile data gathering and energy provisioning in wireless rechargeable sensor networks. IEEE Trans. Mob. Comput. **13**(12), 2836–2852 (2014)
14. Johnson, J., Basha, E., Detweiler, C.: Charge selection algorithms for maximizing sensor network life with UAV-based limited wireless recharging. In: 2013 IEEE Eighth International Conference on Intelligent Sensors, Sensor Networks and Information Processing, pp. 159–164 (2013)
15. Katsidimas, I., Nikoletseas, S.E., Raptis, T.P., Raptopoulos, C.: Efficient algorithms for power maximization in the vector model for wireless energy transfer. In: Proceedings of the 18th International Conference on Distributed Computing and Networking, Hyderabad, India, 5–7 Jan 2017
16. Levin, D.A., Peres, Y., Wilmer, E.L.: Markov Chains and Mixing Times. American Mathematical Society, Providence, R.I. (2009). http://opac.inria.fr/record=b1128575
17. Li, Z., Peng, Y., Zhang, W., Qiao, D.: J-RoC: a joint routing and charging scheme to prolong sensor network lifetime. In: Proceedings of the 2011 19th IEEE International Conference on Network Protocols, ICNP '11, pp. 373–382. IEEE Computer Society, Washington, DC, USA (2011). http://dx.doi.org/10.1109/ICNP.2011.6089076
18. Lu, S., Wu, J., Zhang, S.: Collaborative mobile charging for sensor networks. In: Proceedings of the 2012 IEEE 9th International Conference on Mobile Ad-Hoc and Sensor Systems (MASS), pp. 84–92. IEEE Computer Society, Washington, DC, USA (2012)
19. Luo, J., He, Y.: Geoquorum: load balancing and energy efficient data access in wireless sensor networks. In: INFOCOM, 2011 Proceedings IEEE, pp. 616–620 (2011)
20. Madhja, A., Nikoletseas, S., Raptis, T.P., Raptopoulos, C., Tsolovos, D.: Peer-to-peer wireless energy transfer in populations of very weak mobile nodes. In: 2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW), pp. 1–6 (2017). https://doi.org/10.1109/WCNCW.2017.7919073
21. Madhja, A., Nikoletseas, S., Raptopoulos, C., Tsolovos, D.: Energy aware network formation in peer-to-peer wireless power transfer. In: The 19th ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (MSWiM) (2016)

22. Madhja, A., Nikoletseas, S.E., Raptis, T.P.: Distributed wireless power transfer in sensor networks with multiple mobile chargers. Comput. Netw. (2015)
23. Madhja, A., Nikoletseas, S.E., Raptis, T.P.: Hierarchical, collaborative wireless energy transfer in sensor networks with multiple mobile chargers. Comput. Netw. (2016)
24. Mertzios, G.B., Nikoletseas, S.E., Raptopoulos, C.L., Spirakis, P.G.: Stably computing order statistics with arithmetic population protocols. In: 41st International Symposium on Mathematical Foundations of Computer Science, MFCS 2016, 22–26 Aug 2016, Kraków, Poland, pp. 68:1–68:14 (2016)
25. Michail, O., Spirakis, P.G.: Simple and efficient local codes for distributed stable network construction. In: Proceedings of the 2014 ACM Symposium on Principles of Distributed Computing, PODC '14, pp. 76–85. ACM, New York, NY, USA (2014)
26. Naderi, M., Chowdhury, K., Basagni, S., Heinzelman, W., De, S., Jana, S.: Experimental study of concurrent data and wireless energy transfer for sensor networks. In: 2014 IEEE Global Communications Conference (GLOBECOM), pp. 2543–2549 (2014)
27. Nikoletseas, S., Raptis, T.P., Raptopoulos, C.: Energy balance with peer-to-peer wireless charging. In: 2016 13th IEEE International Conference on Mobile Ad hoc and Sensor Systems (MASS) (2016)
28. Nikoletseas, S., Raptis, T.P., Raptopoulos, C.: Interactive wireless charging for energy balance. In: 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS), pp. 262–270 (2016)
29. Nikoletseas, S., Raptis, T.P., Raptopoulos, C.: Interactive wireless charging for weighted energy balance. In: 2016 International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 119–121 (2016)
30. Nikoletseas, S., Raptis, T.P., Souroulagkas, A., Tsolovos, D.: An experimental evaluation of wireless power transfer protocols in mobile ad hoc networks. In: 2015 IEEE Wireless Power Transfer Conference (WPTC), pp. 1–3 (2015)
31. Nikoletseas, S., Raptis, T.P., Souroulagkas, A., Tsolovos, D.: Wireless power transfer protocols in sensor networks: Experiments and simulations. J. Sens. Actuator Netw. **6**(2) (2017). https://doi.org/10.3390/jsan6020004. http://www.mdpi.com/2224-2708/6/2/4
32. Nikoletseas, S., Yang, Y., Georgiadis, A. (eds.): Wireless Power Transfer Algorithms, Technologies and Applications in Ad Hoc Communication Networks. Springer International Publishing (2016)
33. Peng, Y., Li, Z., Zhang, W., Qiao, D.: Prolonging sensor network lifetime through wireless charging. In: 2010 IEEE 31st Real-Time Systems Symposium (RTSS), pp. 129–139 (2010)
34. del Prete, M., Costanzo, A., Georgiadis, A., Collado, A., Masotti, D., Popovic, Z.: Energy-autonomous bi-directional Wireless Power Transmission (WPT) and energy harvesting circuit. In: 2015 IEEE MTT-S International Microwave Symposium (IMS) (2015)
35. Rault, T., Bouabdallah, A., Challal, Y.: Multi-hop wireless charging optimization in low-power networks. In: IEEE Global Communications Conference (GLOBECOM), pp. 462–467 (2013)
36. Richa, A.W., Mitzenmacher, M., Sitaraman, R.: The power of two random choices: a survey of techniques and results. Comb. Optim. (2001)
37. Rolim, J.: Energy balance mechanisms and lifetime optimization of wireless networks. In: Contemporary Computing. Communications in Computer and Information Science, vol. 168. Springer, Berlin, Heidelberg (2011)
38. Schafer, S., Coffey, M., Popovic, Z.: X-band wireless power transfer with two-stage high-efficiency GaN PA/rectifier. In: IEEE Wireless Power Transfer Conference (WPTC) (2015)
39. Sobin, C., Raychoudhury, V., Marfia, G., Singla, A.: A survey of routing and data dissemination in delay tolerant networks. J. Netw. Comput. Appl. **67**, 128–146 (2016). https://doi.org/10.1016/j.jnca.2016.01.002
40. Ulukus, S., Yener, A., Erkip, E., Simeone, O., Zorzi, M., Grover, P., Huang, K.: Energy harvesting wireless communications: a review of recent advances. IEEE J. Sel. Areas Commun. **33**(3), 360–381 (2015)
41. Wang, C., Li, J., Ye, F., Yang, Y.: Multi-vehicle coordination for wireless energy replenishment in sensor networks. In: 2013 IEEE 27th International Symposium on Parallel and Distributed Processing, pp. 1101–1111 (2013)

42. Wang, C., Li, J., Ye, F., Yang, Y.: NETWRAP: an NDN based real-timewireless recharging framework for wireless sensor networks. IEEE Trans. Mob. Comput. **13**(6) (2014)
43. Wang, C., Li, J., Ye, F., Yang, Y.: Recharging schedules for wireless sensor networks with vehicle movement costs and capacity constraints. In: 2014 Eleventh Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), pp. 468–476 (2014)
44. Xiang, L., Luo, J., Han, K., Shi, G.: Fueling wireless networks perpetually: a case of multi-hop wireless power distribution. In: 2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp. 1994–1999 (2013)
45. Xie, L., Shi, Y., Hou, Y.T., Sherali, H.D.: Making sensor networks immortal: an energy-renewal approach with wireless power transfer. IEEE/ACM Trans. Netw. **20**(6), 1748–1761 (2012)
46. Zhao, M., Li, J., Yang, Y.: A framework of joint mobile energy replenishment and data gathering in wireless rechargeable sensor networks. IEEE Trans. Mob. Comput. (2014)

Next-Generation Software-Defined Wireless Charging System

M. Yousof Naderi, Ufuk Muncuk and Kaushik R. Chowdhury

Abstract Recent research in the emerging field of RF wireless energy transfer and harvesting has shortcomings such as low charging rates, and real-time adaptability and intelligent control to changing energy demands of the network. In this chapter, we introduce DeepCharge, a new architecture for next-generation wireless charging systems that act as an integrated hardware and software solution, and consists of software controller, programmable energy transmitters with distributed energy beamforming, and multiband energy harvesting circuits. DeepCharge realizes a software-defined wireless charging system through separation of controller, energy, and hardware planes. We demonstrate our indoor and outdoor prototypes with extensive experimental measurements, and discuss the RF exposure safety limits besides the most important research challenges toward next-generation DeepCharge-based wireless charging architectures and systems.

1 Introduction

Electronic devices, from implants to IoT sensors, are being increasingly integrated into our daily life in a wide variety of applications. Industry predictions state that 50 billion devices will be connected to the Internet in the next 5 years [1, 2]. Energy continues to be a key challenge in these devices. On one hand, they have limited battery, while on the other hand, there is a need for consumer-friendly and fast recharging methods. Wireless charging can potentially realize battery-less Internet

M. Yousof Naderi (✉) · U. Muncuk · K. R. Chowdhury
Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA
e-mail: naderi@coe.neu.edu

U. Muncuk
e-mail: umuncuk@coe.neu.edu

K. R. Chowdhury
e-mail: krc@ece.neu.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_15

of things (IoT) and eliminate the need for external power cables or periodic battery replacements.

Some of the shortcomings of state-of-the-art wireless charging are: (i) need for direct contact between wireless charger and receiver during inductive-based charging that imposes location constraints and less freedom, (ii) low charging speed, (iii) low charging ranges, and (iv) lack of a distributed system, intelligent controller, and real-time adaptability to energy demands.

In this book chapter, we introduce our vision of a software-defined wireless charging system called DeepCharge, for intelligent charging. The proposed architecture relies on a combination of controllable software and hardware. Specifically, programmable energy transmitters (ETs) and energy harvesters (EHs) are the main components of a network architecture that provides wireless charging without the need for a fixed power source or charging cable. This design can enable high charging rates and untether electronic devices from the constraints of cables and power sockets, leading to flexibility of deployment and ease of maintenance. In this chapter, we present the system architecture, the system prototypes for both indoor and outdoor scenarios, extensive experimental studies, research opportunities and challenges, and future directions.

DeepCharge allows carefully controlled energy beamforming from distributed dedicated ETs to increase the charging ranges and capacity of wireless energy transfer. While this form of network driven energy delivery is highly controlled, it also utilizes the RF ambient power (mainly applicable in the outdoor scenarios) to power the IoT devices whenever possible. The growing interest and demands on mobile services have resulted in a constant increase of the installed base stations (BSs) worldwide. It is predicted that cellular coverages will increase to over 95, 90, and 65% of the world population, respectively, by 2019 [3]. With such rapid scaling of communication infrastructures, cellular and TV signals may soon become important sources of pervasive ambient energy. This can also become a game-changing localized energy source as base stations become smaller and start to be deployed in high densities in both urban and rural areas. While ambient energy harvesting has the advantage of scavenging existing pervasive radiation without any need of dedicated transmitter, how much energy can be effectively delivered is subject to the characteristics of the surrounding environments, schedule followed by the base stations and mobile users, underlying dynamic channel characteristics, line of sight and blockages, and distance to the ambient sources. All these factors are considered in the adaptive resource management feature integrated within DeepCharge.

The contributions of the proposed DeepCharge network architecture are threefold: First, we enable energy transfer through the distributed action of multiple ETs emitting radio frequency (RF) radiation. We demonstrate prototypes for both indoor and outdoor scenarios, accompanied by extensive experimental studies. Second, we present the evolution of our RF energy harvesting circuits leading to a design that can operate at very low levels of input RF signal strengths for ambient harvesting and also high power levels. Our circuit not only harvests energy intentionally transmitted in the license-free industrial, scientific, and medical (ISM) 2.4 GHz and 900 MHz band, also used by Wi-Fi, Bluetooth, by dedicated energy transmitters, but can also

scavenge ambient energy present in the cellular GSM and LTE bands. Third, we develop a software controller that schedules multiuser wireless charging and adapts optimal sequence and duration of energy beams to the energy needs. In particular, the software platform remotely manages the hardware profiles, resources (e.g., energy waveforms and transmission powers), and coordinates the actions of multiple energy transmitters so that the net result in a focused energy targeting a particular sensor.

The DeepCharge architecture offers higher levels of adaptation and reconfigurability for wireless charging through software-defined architecture with controller software and programmable ETs, and higher wireless charging rates and distances through spatially distributed wireless beamforming compared to omni-directional RF energy transfer. The DeepCharge system provides convenience regarding the charging IoT devices with varying and high demanding charging needs and impacts scenarios where city-wide deployed sensors may not have access to sunlight and are impaired by dust/snow accumulation on solar panels. In such cases, sensors powered by ambient cellular transmissions would be critical.

The rest of this chapter is organized as follows. Section 2 gives background information and a summary of the most relevant works. This is followed by a description of the DeepCharge architecture in Sect. 3. Our indoor and outdoor implementations as well as the evolution of our energy harvesting circuits are detailed in Sect. 4. Section 5 discusses the RF exposure safety and FCC regulations, and Sect. 6 provides a summary of emerging research challenges for software-defined wireless charging systems. Finally, Sect. 7 concludes our work.

2 Related Works

2.1 Energy Harvesting-Powered IoT

First, we discuss important works regarding traditional energy harvesting-powered IoT systems. Solar is one of the earliest candidates for energy harvesting in sensor networks, since it offers high energy density among the ambient options, and is available in a wide range of sizes and power levels [4, 5]. In [6], a solar energy harvesting module is used to establish a solar-powered network where some nodes can receive and transmit packets without the need to consume their limited battery resources. In [7], a solar energy harvesting system based on the principle of photo-voltaic effect is proposed. In addition, a new recharging circuitry, which can reinforce the lifetime of the nodes, is designed to recharge its battery when the charge drops below a threshold level. Brunelli et al. [8] utilizes the automatic maximum power point tracking at a minimum energy cost by minimizing the size of harvesters photo-voltaic modules. The proposed harvesters power consumption is less than 1 mW. In [9], the authors proposed a new low-power maximum power point tracker (MPPT) circuitry for sensor network, which transfers the energy in nonoptimal weather conditions. The integrated solar energy harvester presented in [10], scavenges the energy using

an array of photo diodes, fabricated with storage capacitors on a chip. Despite the benefits of solar harvesting, it has a major drawback of operation only in the presence of direct sunlight.

Another practical source of energy harvesting is mechanical vibration. A vibration energy harvester, presented in [12], scavenges electricity from the force exerted on shoes during walking using piezoelectric crystals. An example demonstration of vibration energy harvesting from the ambient environment is proposed in [13] to drive an autonomous wireless condition monitoring sensor system (ACMS). This system generates average 58 W in a volume of only 150 mm^3 at 52 MHz, when the system is used with ACMS. This system has commercialized as an industrial air compressor and an office air conditioning unit. In [14], the aim is to achieve higher than 10 W per day when the person wears the watch and moves the hand. The important aspect for vibration energy harvesting is the size of the generator. [15] presents a mechanism to gather energy based on motion-driven generator with linear motion of the proof mass. According to the experiments, with 0.25 mm^3 generator, the device is able to generate power between 1 and 4 W. With an 8cm3 generator, the power between 0.5 and 1.5 mW is measured as an output.

Wearable devices can be powered with human body energy harvesting. One of the approaches is to power devices from human body heat. According to [16], the first thermo-electric device utilizing this concept was demonstrated in 2004. Interuniversity Microelectronics Center (IMC) fabricated a watch-size thermo-electric generator, which can scavenge energy using human body heat to generate about 100 W under normal activity. Moreover, IMEC has developed a wearable (headphone type) battery-less wireless two-channel electroencephalography (EEG) system, which is powered by heat and ambient light. This system can generate more than 1 mW, on average, in an indoor environment while consumes only 0.8 mW [17]. Recently, the research community has started to focus on improving efficiency of these thermo-electric devices. In a recent study presented in [18, 19], thermal impedance matching as well as electrical impedance matching are two important aspects of improving the performance of thermo-electric devices affixed to the human body.

Wireless energy transfer for powering IoT devices has received significant attention in the recent years, with comprehensive classification and surveys on this topic presented in [22–24, 83]. However, the concept of wireless energy transfer is not new and was proposed first by the Nikola Tesla back in 1899. Based on Teslas research, W.C. Brown introduced and developed the first rectenna, which is a type of voltage rectifying antenna in the 1960s. The first experimental results on the rectenna in 1963 reveal an efficiency of 50% and output 4WDC and 40% at output 7WDC using 23 GHz as the operational frequency. Additionally, devices were developed based on microwave power transmission (MPT) from 1964 to 1975, such as using MPT in helicopters and design of further refined rectennas whose efficiencies ranged from 26.5% at 39WDC to 54% at 495WDC with amplitron, an oscillator (or a generator of microwaves) that could amplify a broad band of microwave frequencies, at 2.45 GHz [20]. Many researchers recently have focused attention on RF energy harvesting, despite its low energy density. A wireless battery charging system using RF energy harvesting was studied in [25]. In the study, a cellular phone can be charged

with the charging rate of 4 mV/s at a frequency of 915 MHz. RF energy harvesting with ambient sources is presented in [26], where the energy harvester can obtain 109 W at 800 MHz from daily office workers' routine in Tokyo. A new design for remote telemetry based on RF energy harvesting was proposed in [21]. The design is capable of generating and delivering RF energy for down-hole telemetry systems that is used for monitoring levels of underground water and fossil fuel sources, using conductive pipes radiating RF signal. Consequently, the down-hole telemetry systems may be converted to wireless systems with the help of this design.

Among the recent works, [72, 73] design and develop a communication system for data over energy transfer, and [75, 79] introduce techniques for multi-hop energy transfer. In the case of in-band energy and data transfer, [76] provides a routing protocol and [80, 85, 86] design medium access controls that address challenges of how and when should the energy transfer occur. These works address research challenges at the protocol level such as the impact of energy communication over data, when schedule energy transfer, decide and assign right priority for energy transfer, and allocate optimal frequencies that maximize received power. Thus, energy transfer becomes a complex medium access problem that needs to address additional parameters such as battery level, wave propagation effects, and energy harvesting efficiency characteristics in addition to the classical link layer problem that aims to achieve fairness in the channel access. Furthermore, the act of multiple concurrent energy transfers introduces both constructive and destructive interferences among RF waves from different ETs. Being able to compute the harvestable energy at a given point in space is nontrivial, and depends on the relative distance of active ETs as well as path loss information. Analytical frameworks have been developed for sensor networks with multiple ETs to model the distributions of wireless charging times [82], node lifetime [81], and harvestable RF energy [78]. In addition, RF-powered IoT networks with concurrent data and energy transfer [74, 84] and mobile chargers [87] have been experimentally studied.

Common areas where RF energy harvesting concept is utilized includes passive radio frequency identification (RFID) and passive RF tags. RFID and RF tags contain a device powered by propagating RF waves [27, 28]. In [29], the authors investigate the feasibility and potential benefits of using passive RFID as a wake-up radio. The results show that using a passive RFID wake-up radio offers significant energy efficiency benefits at the expense of delay and the additional low-cost RFID hardware. Recently, prototypes for such RF harvesters have been developed in both academia [30, 31], as well as commercial products in the industry [32].

2.2 Distributed Wireless Energy Transfer

Early wireless charging efforts mainly involves single-antenna omni-directional energy transfer and multi-antenna beamforming with a single transmitter/base station. However, nowadays, many communication systems are equipped with more than one transmitter, and also the need for distributed beamforming is emerging. Yang

et al. [55] has introduced an adaptive energy beamforming scheme using imperfect channel state information (CSI) feedback in a point-to-point multiple-input single-output (MISO) system. It maximizes the harvested energy while taking into consideration balancing the time resource used for wireless power transfer and channel estimation. Additionally, [56, 57] have studied energy beamforming in multiuser systems. Sun et al. [56] considered a TDMA-based MISO system and modeled a joint time allocation and energy beamforming design as a non-convex programming problem to maximize the system sum-throughput. Moreover, [58] has introduced a novel signal splitting scheme at the transmitters that optimize the rate-energy trade-off and exploits collaborative energy beamforming with distributed single-antenna transmitters.

2.3 RF Energy Harvesting Circuits

We have seen recently interesting developments in higher efficient energy harvesting for dedicated and ambient RF source with a diverse range of input powers from low to high. Table 1 summarizes and compares some notable works.

Park et al. introduced an RF energy harvesting design with an efficiency of 78% at more than 10 dBm using a dedicated RF source [37]. Scorcioni et al. [38] designed a circuit with 40% efficiency at inputs higher than −10 dBm. Le et al. [39] integrated a CMOS-based RF-DC power converter to obtain 60% efficiency at −8 dBm. On the other hand, [40] proposed a dual-stage design to operate both in the high input power and the low input power ranges up to 20 dBm. Furthermore, ambient RF energy harvesting circuit designs from sources such as Wi-Fi, cellular base stations, and analog/digital TV have been advanced in the form of single circuit and array of circuits. Nishimoto et al. [44] has demonstrated a sensor node that harvests from TV transmitters. Vyas et al. [45] harvest single-tone digital TV signals at a distance of 6.3 km from the source, and showed a circuit with a peak efficiency of 20% at the input power of −3 dBm. Additionally, [47] presented a self-powered sensor at 10.4 km from the source by harvesting digital TV signals.

Harvesting power from multiple frequency bands simultaneously is an emerging concept with some notable works on [43, 48, 50, 51], where multiband array rectennas have multiple antennas tuned to the individual bands. On the other side, [52] showed a multiband rectifier with a wideband antenna, instead of multiple antennas, and a summation network is described in that enables combining power from these rectifiers, even if all bands are not available at the same time in the environment.

2.4 Wireless Software-Defined Networking

Software-defined networking (SDN) defines a new architecture that allows the utilization of one or multiple controllers to manage the resources in a network. The key

Table 1 Comparative evaluation of state-of-the-art RF energy harvesting circuits and systems

Related work	Frequency band	Peak conversion efficiency	Matching architecture	Technology
Keyrouz [35]	DTV (470 – 810 MHz)	NA	L matching network	HSMS-282C HSMS-285C
Parks [36]	DTV (539 MHz) ISM (915 MHz) 267, 400, 600, 900 and 1350 MHz	25 – 30% @ −10 dBm 5 – 10% @ −10 dBm	L Matching Network	HSMS-285C
Scorcioni [38]	ISM (868 MHz)	58%@−3 dBm	NA	130 nm CMOS
Nintanavongsa [40]	ISM (902 – 928 MHz)	NA	NA	HSMS-2852 HSMS-2822
Liu [43]	GSM900 (915 MHz) GSM1800 (1800 MHz)	30%	Transmission line	HSMS-2850 ATF34143
Nishimoto [44]	DTV (512 – 566 MHz)	8% @ 10 KΩ	NA	NA
Vyas [45]	DTV (512 – 566 MHz)	19.5% @ 1 M Ω	L matching network	NA
Shigeta [46]	DTV (512 – 566 MHz)	7% @ 470 K Ω	L matching network	SMS-7630 HSMS-286C
Parks [47]	DTV (539 MHz) Cellular (738 MHz)	23% @ −8.8 dBm 26% @−18 dBm	NA	Seiko S-882Z IC Charge Pump
Powercast 2110B [49]	ISM (902 – 928 MHz)	NA	NA	CMOS
Pinuela [51]	DTV (470 – 610 MHz) GSM900 (925 – 960 MHz) GSM1800 (1805 – 1880 MHz) 3G (2110 – 2170 MHz)	40% @ −25 dBm (end-to-end)	L Matching network	SMS-7630
Stoopman [53]	ISM (915 MHz) and ISM(886 – 908 MHz)	40% @ −17 dBm	NA	90 nm CMOS
Assimonis [54]	ISM (868 MHz)	9% @ 10 K Ω	Transmission line (Microstrip)	HSMS-285C

idea here is to decouple the control decisions from the data plane (e.g., switches and routers), enable remote management of programmable hardwares, and dynamically configure and update the networked devices over their operational lifetime. However,

SDN has been mainly applied in the areas of wired networks [65], data centers [67], and recently, in wireless networks such as underwater [68] and 5G [66].

A detailed survey on wireless software-defined networks can be found in [59–61] where SDN is utilized to address different challenges, such as dynamic interference management, congestion control, load balancing, enhancing scalability and global view of network, and enabling multiple wireless network coexistence. Additionally, there is a symbiotic relationship between SDN and software-defined radios (SDRs). SDRs provide a radio communication system that can be reconfigured and reprogrammed completely by the user. This allows programmable physical and MAC layers, and replaces hardwares such as mixers, filters, amplifiers, modulators/demodulators, and detectors by software programs. Accordingly, SDRs can be a complementary solution to SDN. Macedo et al. [62] and Jagadeesan et al. [63] discuss the role of SDRs within wireless SDN environments.

Furthermore, network programmability as a feature through the integration of SDRs and SDN has been studied in works such as [61, 64]. Moreover, [66] proposes SoftAir, a system for software-defined underwater networks, and discusses the benefits and the essential management tools for underwater SDN-based networks. Finally, an architecture for wireless SDN (called software- defined wireless virtual network (SDWVN)) has been introduced in [69], where the authors propose a system with three layers such as a physical network (L1), virtualization function (L2), and virtual (L3).

Given the many benefits of SDN, DeepCharge is built atop the control model defined by an SDN to realize the next-generation wireless charging paradigm. As we will demonstrate in the sections below, it is a viable method to power IoT networks, and to best of our knowledge is the first work on this area of SDN-based energy transfer.

3 DeepCharge Architecture Design

As shown in Fig. 1, DeepCharge consists of IoT devices (i.e., sensor nodes, wearable devices, etc.), multiple energy transmitters, and ambient sources such as TV and cellular base stations. There are two sources of energy to power the IoT devices. Intelligent beams from a set of ETs and ambient energy from cellular/TV base stations. The ETs jointly forms a highly focused power in the form of energy pulse and target it to a device for a determined duration. Each sensor is equipped with an energy harvesting circuit which is capable of harvesting energy at different frequencies from ambient and RF beams and converts them to DC power for device operation. A new device can register/authenticate itself to the server controller before receiving the energy. The RF harvesting circuit is tunable to allow harvesting from frequencies in ET and ambient bands. DeepCharge allows ETs to emulate a software-defined and scalable virtual multiple-input multiple-output (MIMO) system that can transmit N concurrent streams to one or more devices that may give an improvement up to N^2 in the gain of the received power compare to noncooperative single ET action.

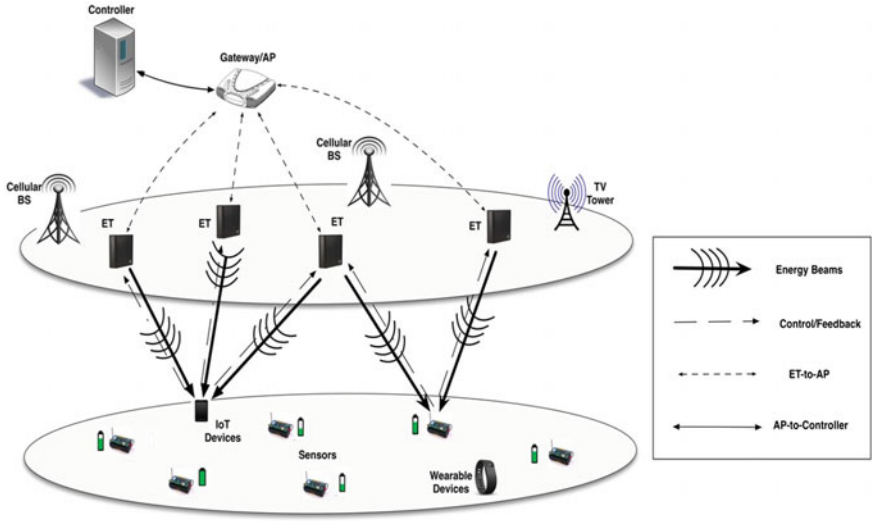


Fig. 1 Overview of DeepCharge architecture

Figure 2 depicts the architecture and interactions of DeepCharge that consists of three layers: controller, software-defined energy and data, and hardware. Distributed wireless charging behavior can more effectively be controlled in the centralized server

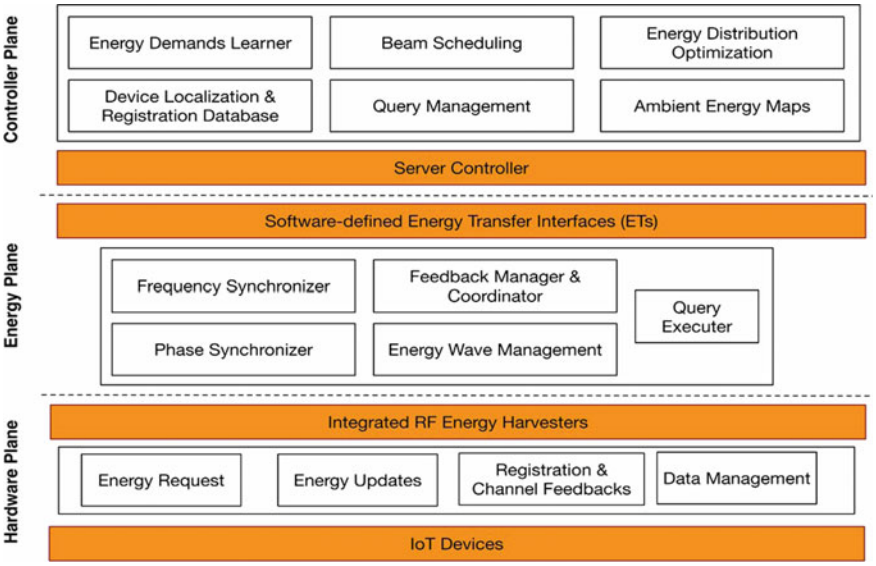


Fig. 2 The layered architecture and interactions of DeepCharge

rather than using custom configurations in different devices scattered across the network in order to mitigate interference and optimize performance of energy transfer. At this layer, through management commands generated by query management module, the server can configure network elements and transmission parameters, enable self-healing capabilities by detecting and controlling energy interference, power outage, and faulty elements, and manage data and energy duty-cycle durations.

To this end, "device localization and registration database" module keeps the records of active devices and their locations, and "energy demands learner" estimates the upcoming energy consumptions based on active data traffic rates, and history of energy demands. Additionally, "energy distribution optimization" module acts for optimizing energy transfer flows over the network while satisfying the estimated energy needs, and "beam scheduling" module handles time scheduling of the beams accordingly.

In the second layer, software-defined energy interfaces are utilized by the server to implement the different behavior policies. This interface allows for effective virtualization of the access network and the best dynamic use of available resources. This layer is responsible for managing physical energy beamforming (e.g., beamforming matrix) through energy wave management and feedback manager, frequency and phase synchronizations, unified control of energy and data access, and signal processing tasks. The functionality of modules related to this layer has been discussed in Sect. 4.1. Finally, the hardware layer is responsible for generating the initial end-device registration requests, energy requests when the battery of an IoT end-device goes below a threshold or needs more energy, and energy level updates of the end-device. In addition, it reports feedbacks such as measured received signal power and channel state information (CSI) and sensor local data.

DeepCharge system architecture provides several advantages. First, it supports the abstraction of the energy and data from the control plane, which results in less complexity, and brings more flexibility into charging system. Moreover, it allows intelligent management based on traffic flows by integrating wireless energy transfer with time-varying traffic and energy demands at the server. In particular, devices may have spatially and temporally varying energy needs based on their application-specific requirements and locations. These different energy needs require different energy wave allocations, tailored to each specific device. The controller uses the latest aggregated energy updates and status of devices to schedule energy beam durations to support current needs. The controller can perform energy transfer optimizations for a large number of deployment scenarios, also considering any global interference scenarios between groups of ETs controlled by different users or installed by distinct establishments. The set of target devices that are served at any given time and the schedules of the distributed energy beams may change adaptively, based on the energy demands and energy status of network notified to the software controller.

Therefore, DeepCharge deals with traffic bursts and manages short-term and long-term energy load balancing among ETs as well as IoT devices. Furthermore, it supports dynamic energy beamforming allocation/scheduling and allows implementations such as (i) determine and control who gets energy, when and for how long; (ii) choose the level of granularity of charging; (iii) dynamically adjusts trans-

fer power to cover needed nodes; (4) shut down or lower the power of ETs to save energy. In addition, the majority of the state-of-the-art resource allocation solutions for wireless charging systems as well as wireless software-defined networking is application-dependent and based on meticulously designed heuristics. However, the complexity of heuristics for real-world resource management scales exponentially with the number of devices. In addition, the scalable wireless systems are complex and often impossible to model accurately. For instance, the capacity of energy transfer varies with ET transceiver characteristics, frequencies of operation, interactions with other devices, and interference on shared resources such as channel, network, etc. Accordingly, DeepCharge can utilize collected global view of the system such as the RF environment, system states, and network operational constraints to make decisions directly from experience. To this end, the controller could use a combination of machine learning and multi-objective optimization to provide a viable alternative to human-generated heuristics. The decisions can be diverse such as (i) modifying channel access parameters at the link layer, (ii) switching to a different transmission scheme at the physical layer (iii) switching between MIMO beamforming and distributed beamforming, (iv) scheduling duration of beams and ETs groups, and (vi) changing the role/task assigned to individual ET or a group of ETs. Finally, we believe the proposed architecture is a perfect fit for enabling wireless charging as a service in 5G and D2D communications.

4 System Prototypes

4.1 Distributed Indoor Development

Commercially available wireless charging systems are not distributed and have limitations such as low charging rates, need line-of-sight alignment, and close contact with the device. To address these challenges, we developed our DeepCharge prototype as a proof of the concept, which powers nodes through distributed energy beamforming. Figure 3 shows the setup, and consists of the following components (1) programmable ET, (2) RF energy harvester circuit, and (3) controller software. Our programmable ET is based on Universal Software Radio Peripheral (USRP) [33] B210 connected to a $750-950$ MHz 10 W HPA-850 RF bay power amplifier [89]. The RF energy harvester has been fabricated and connected to a sensor device (e.g., TI EZ430) to convert RF-to-DC with high efficiency [34].

The software plane has been implemented in the USRPs and executes distributed energy beamforming algorithm to synchronize both phase and frequency, and maximizes the received power at the desired receiver using a power amplifier with maximum allowable power [88, 90, 91]. Based on the feedback from the target, ETs create a virtual antenna array and focus their signals toward the node, such that the emitted waveforms add up constructively at the target. Each ET uses extended Kalman filter [92] to continuously correct the frequency offset between its carrier frequency

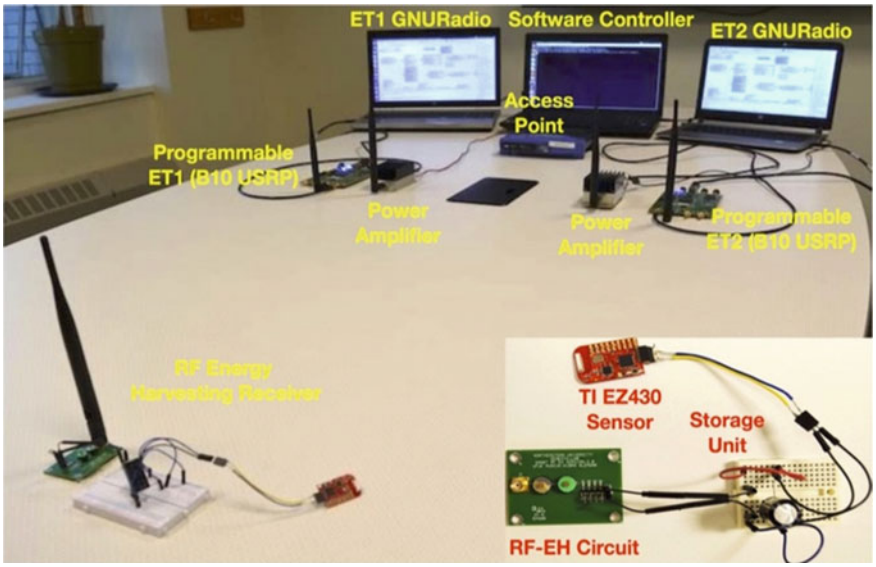


Fig. 3 The experimental setup of DeepCharge for indoor scenario

and the feedback as a reference signal. The energy harvesting circuit converts the incident RF energy into DC voltage stored in the capacitor. The sensor estimates the received signal strength (RSS) of the net incoming signal and broadcasts a single bit to all the ETs to indicate whether this value is higher or lower than that measured in the previous time slot. If the RSS is higher, the ETs update this information and perturb their phase setting using the last setting as the baseline. If the RSS is lower, the ETs revert their phase selection to that of the previous time slot, before beginning the subsequent round of phase perturbation. This randomized ascent procedure is repeated until the ETs converge to phase coherence [93]. Figure 4 shows the results of the energy beamforming for two ETs, and improvements of the received power. There is a synchronization period for phase adjustments among ETs, where the system converges to an optimal value in a few seconds time using the sensor-generated feedback.

The RF energy harvesting circuit contains four components: antenna, impedance matching network sub-circuit, four-stage diode-based Dickson voltage rectifier, and 3300 μF capacitor for energy storage. The impedance matching network sub-circuit with adjustable capacitors maximizes the energy transfer and minimizes power reflection between the antenna and voltage rectifier. For efficient DC conversion, we designed a four-stage diode-based Dickson voltage rectifier by choosing the Schottky diode that operates with quick activation time and lower forwarding voltage drop as the nonlinear component of the rectifier. The 3300 μF capacitor is used to store the energy from the voltage rectifier, which serves as the energy storage for operating the TI EZ430 sensor.

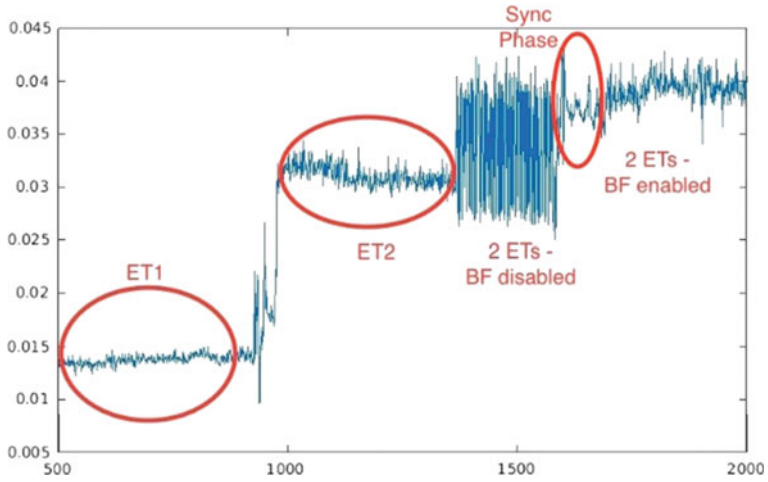


Fig. 4 The received energy comparison according to number of ETs and phase synchronization period

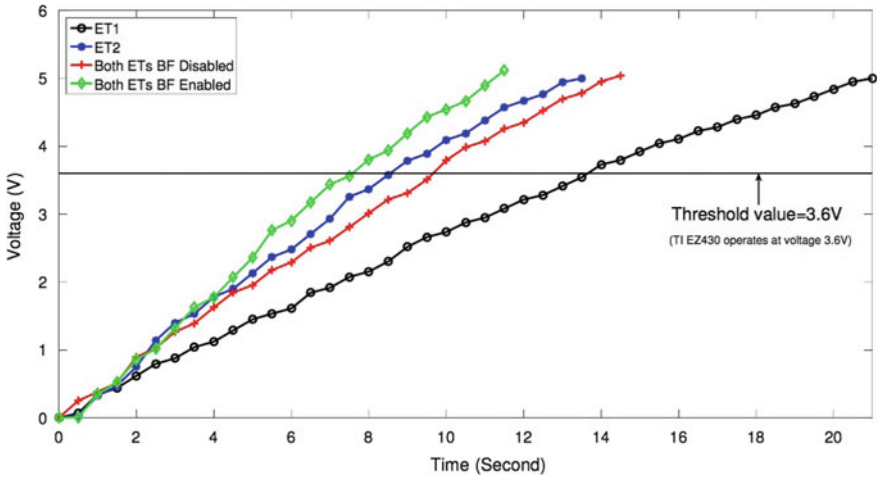


Fig. 5 Comparative charging time for different number of ETs with and without beamforming

As shown in Fig. 5, the net conversion efficiency depends upon the accurate phase matching of the ETs and the circuit design. Once the voltage across the capacitor reaches 3.6 V, the devices disconnect from the charging and resume their normal operation. Furthermore, Fig. 6 shows the instantaneous harvested voltage that has been measured at the output of RF energy harvesting circuit for (a) beamforming-enabled and (b) beamforming-disabled scenarios. It shows the level of harvested voltage when only one ET is turned on, and then two ETs are transmitting. The

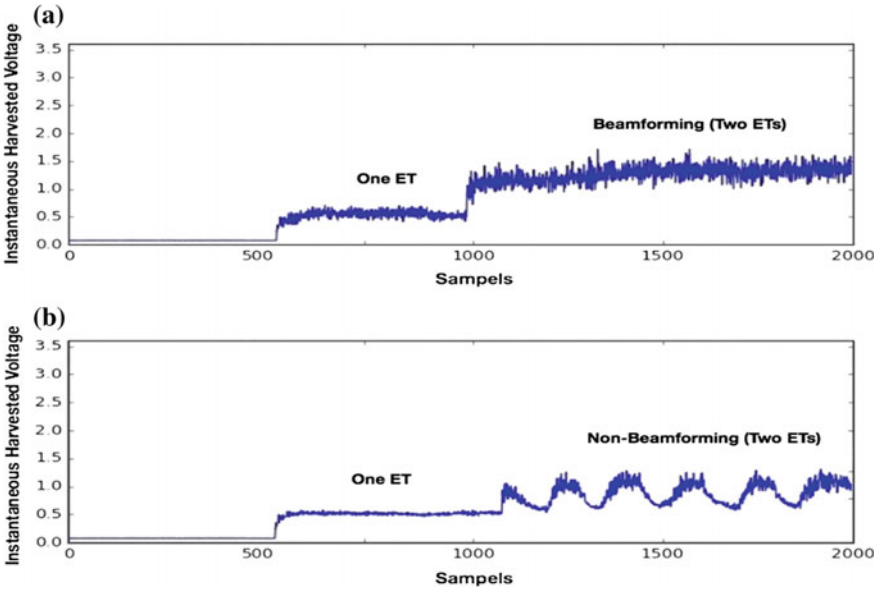


Fig. 6 The instantaneous harvested voltage that has been measured at the output of RF energy harvesting circuit for **a** beamforming-enabled and **b** beamforming-disabled

consistency of constructive interference in the case of our beamforming-enabled implementation is demonstrated.

Finally, our specially designed middleware contains the software controller and communicates with system components (USRPs and receivers) to schedule and manage the hardware resources. Devices and ETs send registration beacons that contain their ID and a description of the functions supported by the discovered device in a predetermined format. The controller then executes remote procedural calls and invokes functions in the ETs, such as adjusting the center frequency, transmit power levels, and stopping or starting beamforming operation, based on the energy status of devices.

The controller implements different resource management and system adjustments such has ET/device registration, adaptive energy allocation, and movement and channel change adaptation. In ET/device registration process, the controller parses all the incoming queries using query management module. If a registration message is detected, controller extracts the request type/ID which refers to either ET or device. Then it registers ET/device with its reported parameters that have been included in the registration message accordingly in a host table (for ETs) or guest table (for devices). The controller then sends appropriate ACK to ET/device to inform successful completion of registration.

Figure 7 shows the process of adaptive energy allocations. The controller constantly monitors the registered devices for their change of energy discharging rates, rates of ambient RF harvesting, and battery levels utilizing the energy updates from

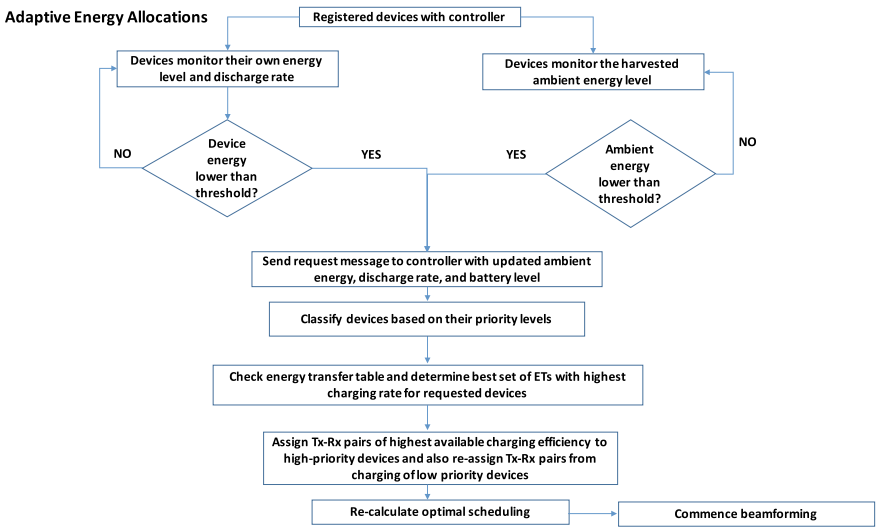


Fig. 7 Depicting the process of adaptive energy allocations based on RF ambient power and discharging rate changes

devices. Each registered device through the RF energy harvester can observe and monitors the level of ambient harvesting rate as well as battery level. When any of these critical parameters goes below specific thresholds an energy request message would be initiated from the device to the controller. Such message contains the updated values of three fields: ambient energy, discharge rate, and battery level. The controller accordingly classifies the devices based on their conditions into different priority levels. It then determines the best set of ETs to serve the highest priority devices by checking the energy transfer table. The energy transfer table contains the end-to-end RF-to-DC conversion efficiency for each set of ETs device. This number is a function of a number of parameters including distance between ETs and receiver device, RF-harvesting circuit, and transmission power. In the initialization phase, the controller estimates these rates using measured harvested powers that are reported from the devices. After updating the best set of ETs and highest priority devices, the controller recalculates the optimal energy beamforming scheduling, and send commands to ETs regarding their schedules/parameters.

In the process of adaptations to any device movement and wireless channel change, each ET receives feedbacks from its target device. These feedbacks can be utilized to estimate the channel and the arrival phase of signals transmitted between the ET device. ET determines any changes in the channel or location change of the energy receiving device through this estimated phase, and update the feedback and steering matrices to provide continuous and accurate energy beamforming. ET informs controller about new changes and the energy transfer table would be updated accordingly.

Table 2 Summary of Boston ambient RF power measurements over 40 subway stations

Band	GSM850	GSM1900	LTE730	LTE740	DTV
Frequencies (MHz)	869–894	1930–1950	734–744	746–756	494–584
Average (mW)	1.8153	0.1335	1.4193	1.4029	0.0547
Maximum (mW)	10.3474	0.5226	13.0601	19.1625	0.3038
Median (mW)	0.7938	0.0821	0.2869	0.0825	0.0362
StDev	2.4500	0.1351	2.9876	3.5793	0.0610

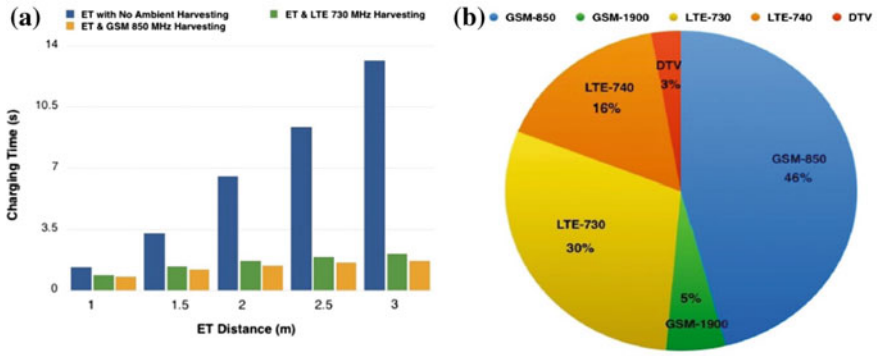


Fig. 8 **a** The charging times according to existence of ambient and beamforming harvesting power, **b** Percentages that each ambient channel wins the highest measured power over all subway stations

4.2 Self-powered Outdoor Development

The importance of outdoor RF energy harvesting is motivated by the recent improvements in RF-to-DC conversion efficiency at low input powers [41], RF outdoor survey studies [42], constant increase of the installed base stations (BSs) worldwide, and increasing advancements in energy efficiency of sensor devices. Next, we have extended DeepCharge system to outdoor scenarios where it utilizes ambient energy harvesting as described in Sect. 4.1. To this end, we first conducted a systematic study of RF energy availability in the Boston city, and then built and optimized our outdoor prototype.

For the first round of experiments, we have conducted a city-wide RF spectral survey at GSM850, GSM1900, LTE730, LTE740, and DTV bands within the Boston area. Our ambient spectrum studies were undertaken from outside of 40 subway stations at street level as survey points that are distributed in the city to measure the available RF power within each ambient band. We used a USRP device with a WBX antenna manufactured by Ettus Research LLC [33], and it was calibrated in the laboratory with the Agilent N9000 signal analyzer. The banded input RF power in mW is calculated by summing and averaging over all received power across

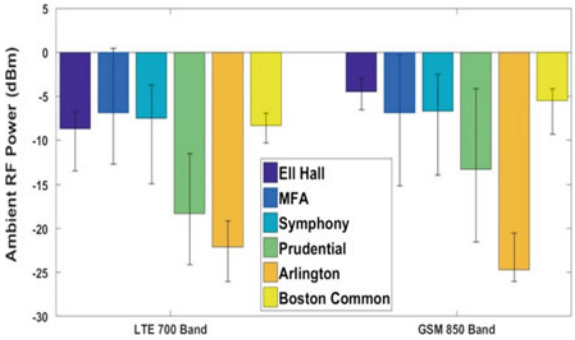


Fig. 9 Levels of average RF power at subway stations, universities, museums, shopping centers, parks in Boston city

the band in a similar way the spectrum analyzer calculates channel power. Table 2 summarizes our results across all subway stations indicating the frequencies, average, maximum, median, and standard deviation for all banded power measurements. It can be observed that RF ambient powers can be relatively high and suitable for harvesting. Based on our survey, Fig. 8a compares the charging times needed to power a Nordic nRF51822 low energy Bluetooth radio for transmission of one packet in three scenarios: (i) ET wireless transfer without any ambient harvesting, (ii) ET wireless transfer with ambient harvesting at LTE 730 MHz, and (iii) ET wireless transfer with ambient harvesting at GSM 850 MHz. Additionally, we found that the percentages of each ambient RF signal band where the highest power is measured is composed of 46, 5, 30, 16, and 3% for GSM850, GSM1900, LTE730, LTE740, and DTV, respectively, as a fraction of all sample locations, as seen in Fig. 8b. This implies that 92% available ambient RF power is contributed by LTE700 and GSM850 bands.

Using insights from our initial survey, follow-up rounds of experiments have been conducted using the RF energy harvester to measure the available ambient RF power at LTE700 and GSM850 bands within six different locations of Boston, such as university areas, museum, shopping centers, outside of the subway station, park, and concert theater. These locations have been selected to help us comprehend the disparities and similarities among results in terms of geographical terrain and population distribution and density when investigating the energy level in such environments. We used our RF harvesting circuit [40] with a PCB Log Periodic antenna, and Sinometer VA18B Multimeter with RS232 USB Cable that is connected to a laptop to record the instantaneous output voltage of the energy harvester. Using Agilent N5181 MXG RF signal generator and Agilent E5061B vector network analyzer, we created a map between output voltages and input powers, and use them to estimate the incident ambient RF power signals.

The measurement samples were obtained with a 30-min section at each location and each section is repeated over the morning of 5 days. The dataset from our survey of the signal strength distribution in Boston is shown in Fig. 9 which summarizes average ambient RF power levels for LTE and GSM frequencies over a set of locations with maximum and minimum values. Accordingly, our proposed RF-EH circuit must

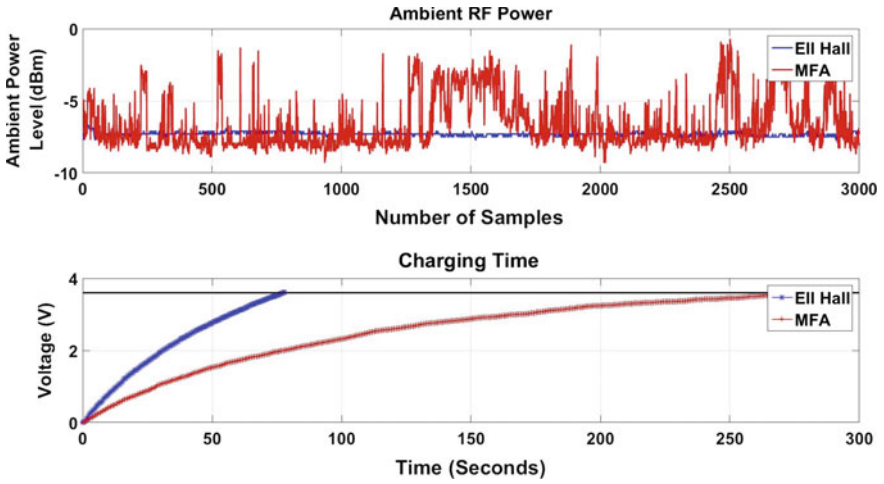


Fig. 10 Comparative illustration of ambient RF power levels and charging times between two locations of Ell Hall and MFA in Boston city

be responsive to a power range between −25 and 0 dBm. The amount of power as well as the fluctuations of ambient RF signals are location-dependent and may vary over time due to physical obstructions in the path, such as multipath fading, attenuation from buildings, reflections, etc. Figure 10 compares ambient power and charging times for two locations: Ell Hall and Museum of Fine Arts (MFA) in Boston. It can be observed that Ell Hall location provides a steady level of power while MFA shows significant fluctuations that have resulted in longer charging time with lower speed.

Figure 11 shows two nodes in our outdoor DeepCharge system prototype, where we have used 6 dBi wideband log-periodic antenna connected to the energy harvester with two TI eZ430-RF2500 master/slave motes with an ultra-low power microcontroller TI MSP430F2274, and 2.4 GHz CC2500 radio chip. Similarity, the experiments were conducted with a 30-min section and repeated over the morning of 5 days. We demonstrate the feasibility of battery-free communications through integration of a two-step charging process within the slave mote: harvest-only and harvest-transmit. We first charge and store ambient power up to 3.6 V, and then follow a duty-cycling mechanism of harvest and data communication. This allows the storage of ambient RF energy in the capacitor by switching between its operational sleep and active modes. During the data communication cycles, the slave sensor periodically wakes up and transmits sensed data, including temperature and voltage values to the master sink and goes back into the sleep state. The master node sends the information to the cloud. In the sleep mode, sensor consumes a current between 7.4 and 15 μA, and in the active mode between 103.7 and 210.3 μA. The slave mote is programmed to operate between 1.8 and 3.6 V to ensure that the sensor cannot enter a nonoperational state. Our design can power a TI eZ430RF2500 sensor continuously in battery-less mode (or active mode) as well as it can harvest and store energy from the ambient RF

Fig. 11 The setup of DeepCharge for self-powered outdoor scenario

power source at the same time. The battery-free operation is shown in Fig. 12 with the charging and data communication (i.e., discharging) duty cycles. It demonstrates the sufficiency of GSM harvested power over charging cycle in front of Ell Hall to provide enough energy for transferring the sensed data over next communication cycle.

4.3 RF Energy Harvesting Circuit Design

RF energy harvesting relies on the energy contained in the RF fields generated by electromagnetic wave transmitters. Conceptually, this energy is captured and converted into functional DC voltage using a specialized circuit directly connected to a receiving antenna. Although this technique has least energy intensity compared to other energy harvesting systems, RF energy harvesting systems have many useful features, not present otherwise. Such systems can be used in any location that has a high incidence of strong ambient RF waves, or in specific applications where there is a presence of a dedicated transmitter. Hence RF energy harvester is generally not dependent on time of the day, geographical aspects of the region, weather conditions etc., which must be considered in other examples of energy harvesting systems

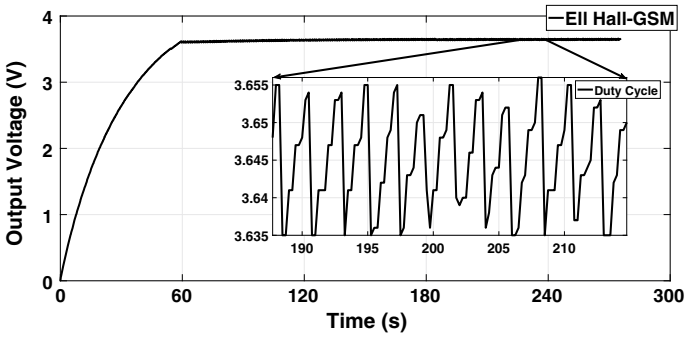


Fig. 12 Depicting charging and data communications cycles with capacitor voltage, when a sensor node periodically wakes up during an outdoor experiment

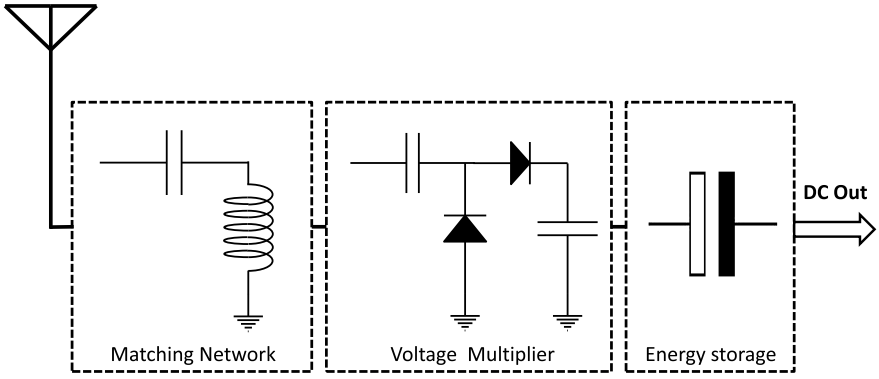


Fig. 13 Overview of RF energy harvesting circuit design

including solar, and wind energy. RF energy can also be used to drive more than one device at the same time.

Figure 13 depicts the components of RF energy harvesting circuit. The voltage multiplier converts incident RF signal into DC power. Impedance matching network consist of inductive and capacitive elements that maximizes the power delivery from the antenna to voltage multiplier. Smooth power load is provided by an energy storage which can be used for durations when external energy is not available. However, a complex interplay of design choices must be considered together. For example, increasing the stages gives higher voltage at the load, but it also reduces the current through the final load branch, and thus result in charging delays. On the other hand, while fewer stages of the multiplier ensure quick charging of the capacitor, the generated voltage might become low and insufficient. Similarly, a slight change in the matching circuit parameters can change significantly the optimal frequency range for harvesting, often by several MHz. Thus, a complex interplay of design

choices must be considered together. In this section, we discuss the most important parameters in our optimal design of RF energy harvesting circuit.

Rectifier Topology: We choose a Dickson topology that provides connected capacitors in parallel, reduces the circuit impedance, and facilitates the matching of antenna side to the circuit load. **Choice of Diodes**: Operating with weak input signals is one of the crucial requirements in the design of RF energy harvesting circuits. Diodes with lowest possible turn-on voltage are preferable since the peak voltage of the AC signal obtained at the antenna is much smaller than the diode threshold [94]. Additionally, diodes with a very fast switching time needs to support harvesting signals at high-frequencies. We use Schottky diodes that have a metal–semiconductor junction and allows the junction to operate much faster and gives a forward voltage drop of as low as 0.15V. **Number of Stages**: The number of rectifier stages has a major influence on the output voltage of the energy harvesting circuit. Each stage is arranged in series as a modified voltage multiplier, and output voltage relates directly to the number of stages. Due to the parasitic effect of the constituent capacitors of each stage, the voltage gain decreases as the number of stages increases. To capture this impact, we conducted measurements using Agilent ADS with sweeping input powers from −20 dBm to 20 dBm and circuit stages from 1 to 9. Figures 14 and 15 show the impact of the number of stages on the efficiency and output voltage of energy harvesting circuit, respectively. The circuit stage in the simulation is a modified voltage multiplier of HSMS-2852, arranged in series. Accordingly, we observe that as the number of stages increases the circuit yields higher efficiency, but for higher stages, the peak of the efficiency curve shifts towards the higher power region. The voltage plot shows that higher voltage can be achieved by increasing the number of circuit stages, but a corresponding increase in power loss is introduced into the low power region.

Impact of Load Impedance: Figure 16 shows the impact of load impedance on five-stage energy harvesting circuit, where each stage is a modified voltage multiplier of HSMS-2852, arranged in series. Here, we have simulated the effect of load impedance on the efficiency using Agilent ADS with a parameter sweep of −20 dBm to 20 dBm and 1 KΩ to 181 KΩ for input RF power and load value, respectively. It can be observed that the circuit yields to optimal efficiency at a particular load value, and its efficiency decreases dramatically if the load value is too low or too high. Over a sensor network operation, the motes draw different amount of current on different operation states: active (all radios operational), low-power (radios shut down for short intervals, but internal micro-controller is active), and deep-sleep (requires external interrupt signal to become active again) states. **Impact of RF Input Power**: RF energy harvesting circuit consists of diodes and shows nonlinear behavior regarding the input power. Accordingly, the impedance of circuit varies with the amount of power received from the antenna. Figure 17 presents the impact of RF input power, ranging from −20 to 20 dBm, on the circuit impedance. A sharp bend at 5 dBm can be observed that is due to this nonlinearity in operation.

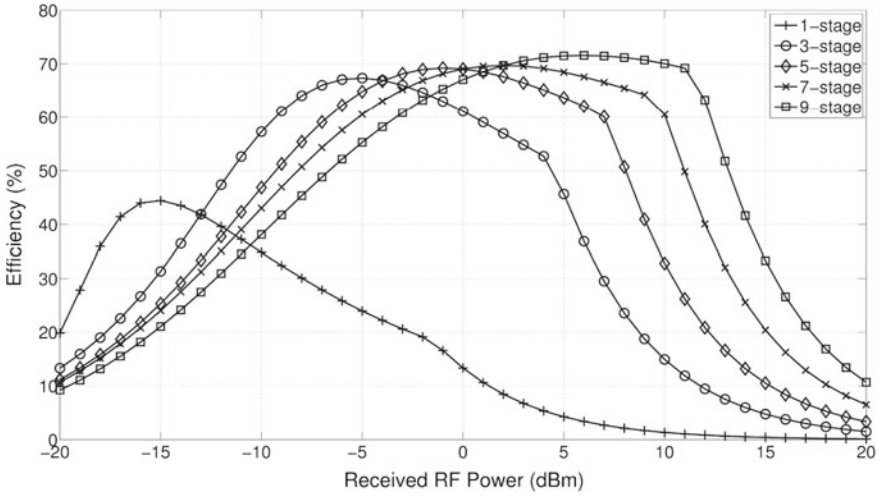


Fig. 14 Effect of number of stages on the efficiency of energy harvesting circuit

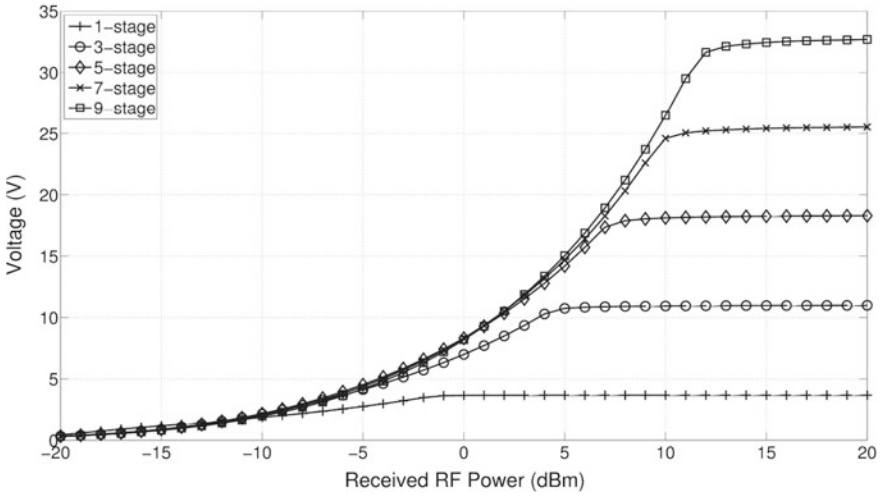


Fig. 15 Effect of number of stages on the output voltage of energy harvesting circuit

4.4 RF Energy Harvesting Circuit Evolutions

In this section, we describe the evolution of our RF energy harvesting circuit in the form of three generation prototypes based on their performance characteristics and usage for different application scenarios.

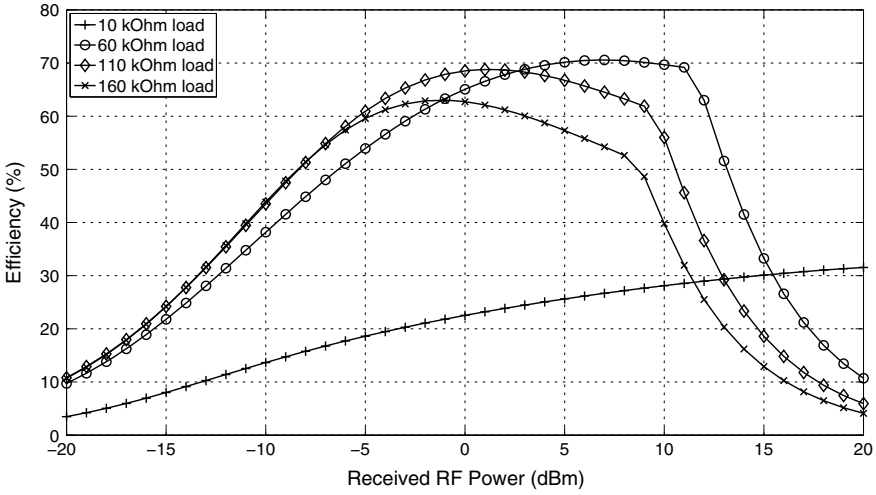


Fig. 16 Effect of load impedance on the efficiency of energy harvesting circuit

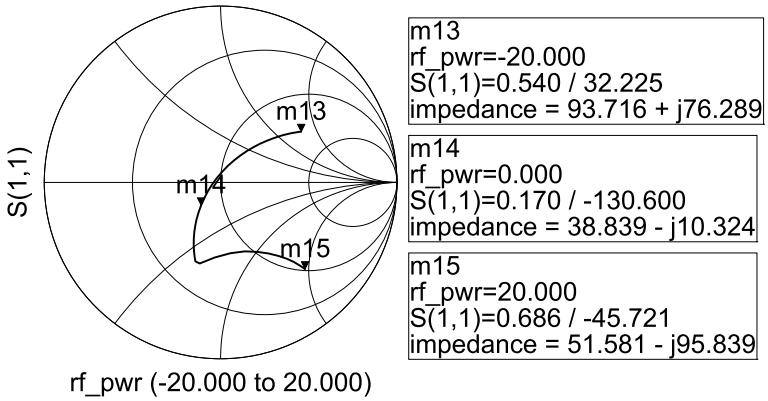


Fig. 17 Effect of RF input power on the impedance of the energy harvesting circuit

4.4.1 First-Generation Design: Dual-Stage RF Energy Harvesting Circuit

Our initial RF-EH prototype [40] consists of a dual-stage energy harvesting circuit with a seven-stage and ten-stage design. Seven-stage is more receptive to the low input power regions and ten-stage is more suitable for higher power range.

Figure 18 shows the first-generation fabricated circuit based on a modified Dickson voltage multiplier. We employed two different diodes from Avago Technologies, HSMS-2822 for high power design (HPD) and HSMS-2852 for low power design (LPD). This circuit can power perpetually Mica2 sensor motes as well as Texas Instruments MSP430G2553 with right duty-cycle configuration. As shown in Fig. 19, the

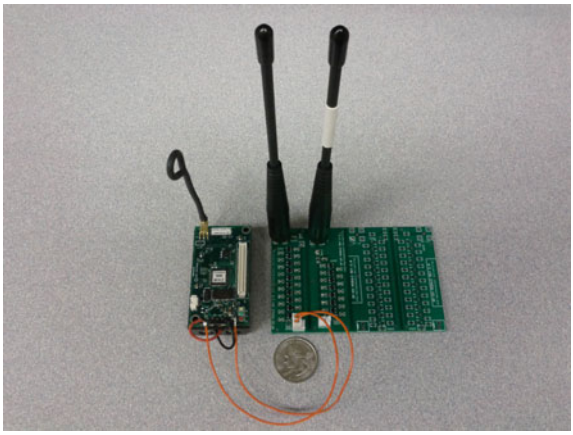


Fig. 18 Overview of our first RF-EH circuit prototype

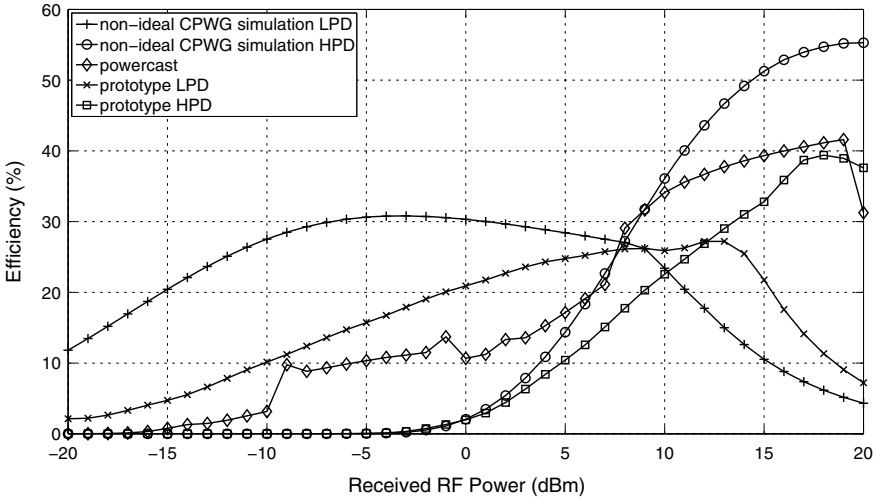


Fig. 19 Efficiency comparison of simulation, prototype, and Powercast energy harvesting circuit

efficiency of this design for both the fabricated circuit and simulated PCB has been compared with the commercial off-the-shelf high-efficient Powercast P1100 circuit. Agilent ADS simulation with coplanar waveguide with the ground plane (CPWG) is used to study the impact of the PCB. Figure 19 shows our prototype and outperforms Powercast P1100 largely between −20 and −7 dBm.

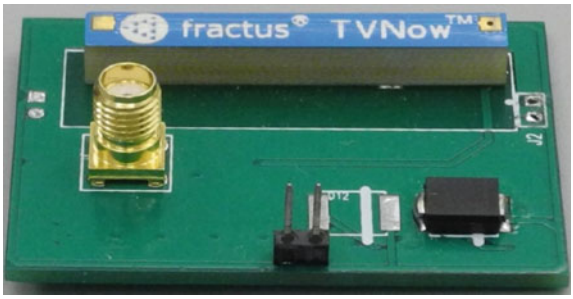
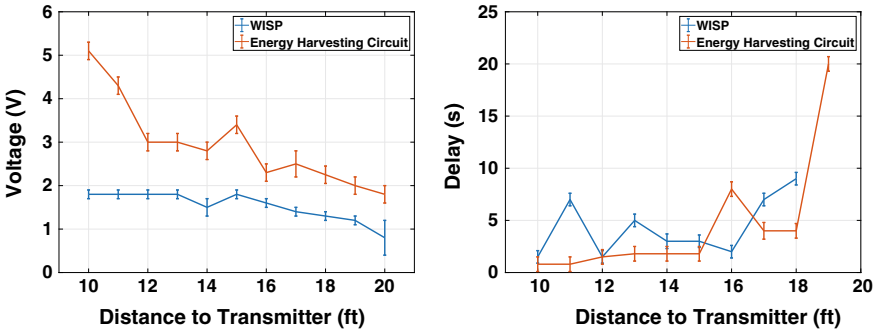


Fig. 20 The second RF-EH circuit prototype as a wake-up radio



(a) Voltage (b) Delay

Fig. 21 Output voltage of RF energy harvesting circuit and potential wake-up delay

4.4.2 Second-Generation Design: RF Energy Harvesting Circuit for Wake-Up Radio

Our second-generation RF-EH circuit prototype [77] is shown in Fig. 20 and acts as a low-cost long-range wake-up receiver. Here, the energy harvesting circuit can generate the pulse signal up reception of a wake-up signal and trigger the interrupt of pin on a connected sensor mote. Figure 21 compares the performance of this circuit with WISP which is an RFID-based battery-free programmable sensing platform. As shown in Fig. 21, our design can get a wake-up response at 19 ft with only 95 s delay and helps to reduce potential wake-up delay and performs much better than WISP.

Additionally, our second RF-EH circuit prototype has been adapted [70] as cost-effective and long-range passive wake-up receiver for both range-based wake-up (RW) and directed wake-up (DW) as seen in Fig. 22 . Here, it interfaces Texas Instruments eZ430-RF2500 sensor board and achieves a wake-up range up to 1.16 m with +13 dBm transmit power. Furthermore, our empirical study shows that at +30 dBm transmit power, the wake-up distance of the developed RW module is >9 m. High accuracy of DW is demonstrated by sending a 5-bit ID from a transmitter at a bit rate up to 33.33 kbps. It has been shown that using a 3 Watts EIRP energy transmitter,

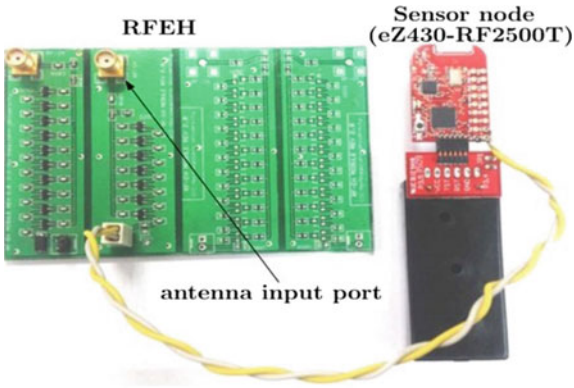


Fig. 22 The second RF-EH circuit prototype for RW and DW passive wake-up radio

compared to a non-optimized design, an optimized design can increase the wake-up range by up to 5.69 times.

Moreover, we have developed an independent wake-up receiver design for DW [71] that is optimized for higher range as well as higher energy savings at both wake-up transmitter and wake-up receiver. By balancing the trade-offs, our optimized design can provide energy saving per bit of about 0.41 mJ and 21.40 nJ, respectively, at the wake-up transmitter and wake-up receiver.

4.4.3 Third-Generation Design: Adjustable Ambient Rf Energy Harvesting Circuit

The range of frequencies for harvesting ambient RF signals, e.g., digital/analog TV and cellular is wide. In the case of array, circuits fabricating more than one circuit that has been pre-tuned to distinct frequency does not allow addition or change of bands after the circuit fabrication, and lacks adaptability to change in the RF input powers. To address these issues, we propose a tunable single energy harvesting circuit [11] that enables harvesting from a wide range of ambient RF signals at LTE 700, GSM 850, and ISM 900 bands. Figure 23 shows the component of our circuit.

Our design enables free choice of any antenna with different characteristic impedances, interfacing between antenna and rectifier, and powering of different sensors regardless of load. This means independent of voltage rectifier from the impedance of both antenna and load. It is a modified π tunable impedance network, and allows ambient RF-EH circuit to select the excited frequency band based on the ambient power.

Figure 24 shows the power conversion efficiency (PCE) performance of our design when connected to a TI eZ430 sensor mote. It can be observed that our circuit can yield up to 48% of performance at sensor active state and more than 20% in the passive state at specified frequency bands.

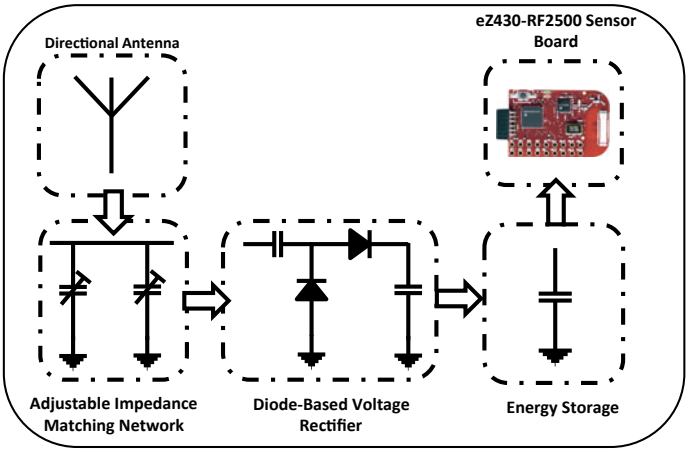


Fig. 23 Overview of our adjustable ambient RF energy harvesting system

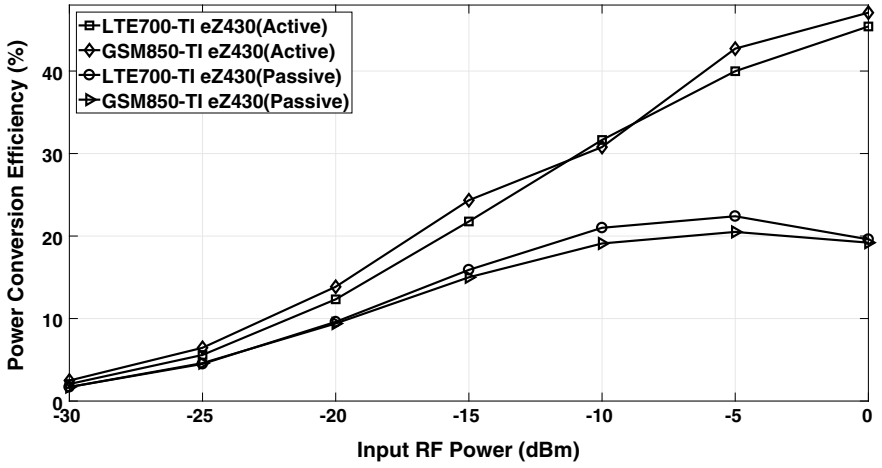


Fig. 24 Performance of our adjustable ambient RF energy harvesting system at LTE 700 and GSM 850 bands in terms of TI eZ430 sensor's operation state

5 RF Exposure Safety and Scalability Analysis

While RF wireless energy transfer technology is an emerging solution to address issues related to battery, reliability, and energy management in next-generation IoT networks, it is paramount to carefully design wireless charging systems that produce RF power below the safety limits of the human body.

The proposed DeepCharge architecture is a promising tool to control and intelligently manage all levels of RF exposures over small and large-scale networks. As per FCC regulations [88, 90, 91], the maximum permissible RF exposure limits for

Table 3 FCC limits for maximum permissible exposure

Frequency range (MHz)	Electric field strength (V/m)	Magnetic field strength (A/m)	Power density (mW/cm^2)	Averaging time (min)
0.3–3.0	614	1.63	100[1]	6
3.0–30	1842/f	4.89/f	$900/f^{21}$	6
30–300	61.4	0.163	1.0	6
300–1500	–	–	f/300	6
1500–100,000	–	–	5	6

Table 4 FCC limits for general population/uncontrolled exposure

Frequency range (MHz)	Electric field strength (V/m)	Magnetic field strength (A/m)	Power density (mW/cm^2)	Averaging time (min)
0.3–3.0	614	1.63	100[1]	30
3.0–30	842/f	2.19/f	$180/f^{21}$	30
30–300	27.5	0.073	0.2	30
300–1500	–	–	f/1500	30
1500–100,000	–	–	1	30

different frequency range are given in the below tables. Different organizations outside US provide similar regulations such as the European Telecommunications Standards Institute (ETSI) for Europe, and Ministry of Posts and Telecommunications (MPT) in Japan (Tables 3 and 4).

Here, the[1] maximum permissible RF exposure limit is expressed in terms of power density mW/cm^2 for certain average time duration. As an example, for the general population (refer table above) in 900 MHz frequency band, the power density limit for 30 min average time is $\frac{f}{1500} = 0.6\,mW/cm^2$, whereas it is $1\,mW/cm^2$ for 2.4 GHz frequency band for the same 30 min average time duration. Next, we show some detailed analysis to manage RF exposures based on FCC regulations. The power density P_d at a distance d can be determined as follows:

$$P_d = \frac{P_T G_T}{4\pi d^2} \tag{1}$$

Assuming all ETs transmit with maximum power P_T, the received signal after will have a beamforming gain up to K^2, where K is the number of participating ETs. Considering the FCC safety regulations, we can calculate the maximum number of ETs which can participate in beamforming as below:

[1] Plane-wave equivalent power density, which has both E-field and H-field components. Equivalent for far-field and near-field power density can be calculated based on: $|E_{total}|^2/3770\,mW/cm^2$, and $|H_{total}|^2/37.7\,mW/cm^2$.

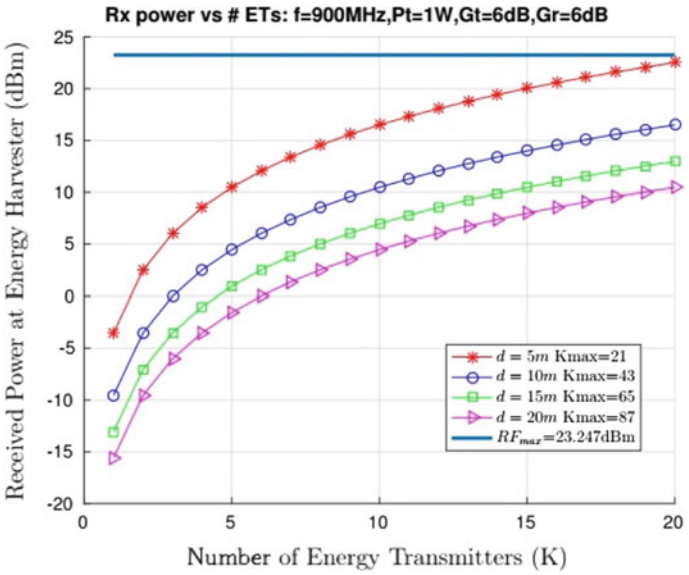


Fig. 25 The numerical plot of received power at energy harvester (in dBm) versus number of ETs for different separation distances

$$K_{max} = \left\lfloor \sqrt{\frac{P_{d,lim}}{P_d}} \right\rfloor \tag{2}$$

For example, if $P_T = 1$ Watt, $G_T = 6$ dB and $d = 1$ m, the power density $P_d = 0.032\,\text{mW/cm}^2$. The FCC RF exposure limit for 900 MHz is $P_{d,lim} = 0.6\,\text{mW/cm}^2$. In this scenario, maximum number of ETs which can participate in beamforming is:

$$K_{max} = \left\lfloor \sqrt{\frac{0.6}{0.032}} \right\rfloor = 4 \tag{3}$$

The maximum allowed exposed RF power can be calculated using free space path loss and K_{max} as below:

$$R_{max} = P_T + G_T + G_R + 20log_{10}(\frac{\lambda}{4\pi d}) + 20log_{10}(K_{max}) = 23.24\,dBm \tag{4}$$

Figure 25 illustrates the numerical plot of the received power at energy harvester (in dBm) versus number of ETs for different separation distances. The blue horizontal line shows the maximum permissible received power according to FCC regulations, and will dictate the maximum number of ETs, which can participate in the wireless charging.

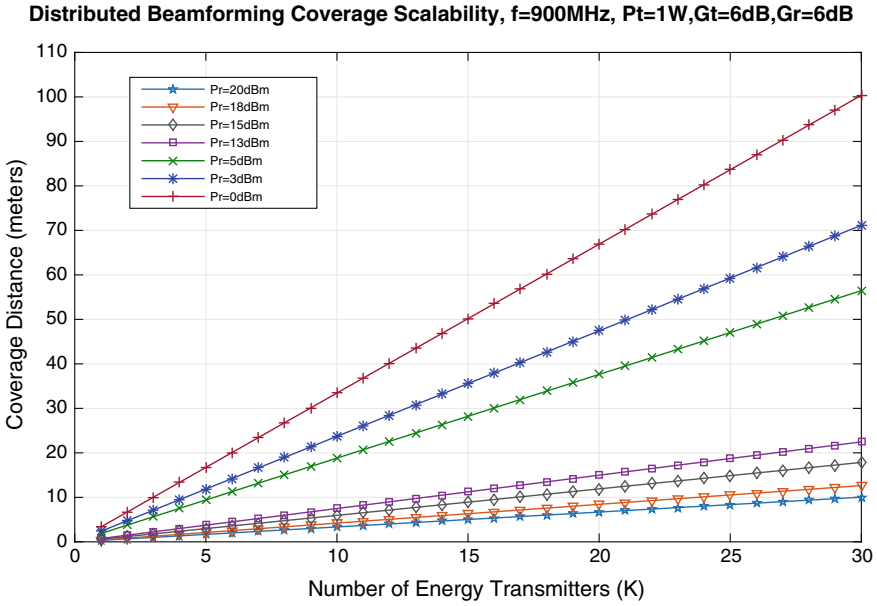


Fig. 26 The coverage distance of distributed beamforming versus number of ETs to guarantee different received powers at the input of RF-EH

Next, we discuss the scalability of system and its constraints. DeepCharge architecture provides a high degree of scalability due to use of distributed beamforming along software controller. Despite MIMO beamforming where multiple antennas attached to a single device and provides a limited coverage and scale, the coverage of DeepCharge increases dynamically by activation or addition of new ETs. In particular, K active ETs results in an increase of gain up to K^2 on the amount of received power in comparison with a single energy transmitter and gain up to $(K)^2 - (K-1)^2$ compare to $K-1$ ETs, which directly translates to increase of charging coverage. The main constraints in scalability here are as follows: (1) convergence time due to phase adjustments of beamforming can increase by the addition of new ETs, (2) maximum permissible RF exposure limits how many active ETs can operate at the same time to guarantee safety regulations, and (3) higher number of ETs could result in higher energy consumptions due to increase in channel estimations.

Figure 26 depicts coverage scalability of distributed beamforming for different number of ETs and received powers (20, 18, 15,13, 5, 3, and 0 dBm) at the input of RF-EH which are enough to power TI eZ430-RF2500 motes with ultra-low power microcontroller TI MSP430F2274, and 2.4 GHz CC2500 radio chip. Each ET transfers with 1 Watts at $f = 900$ MHz and has linear gain 6 dB at both transmit and receive sides. Figure 26 shows how the system can scale by increasing the number of ETs and cover longer charging distances. The convergence time for beamforming optimal phase varies based on wireless channel conditions. For a set of experiments in our

indoor testbed (Sect. 4.1), the average time has been measured 2.4 and 2.9 s for three and five energy transmitters, respectively.

6 Research Challenges

In this section, we discuss the research challenges toward next-generation DeepCharge-based architectures and systems.

1. ***Scalable Coordination of ETs***: Distributed coordination among energy transmitters through time, frequency, and phase synchronizations are one of the key enablers of energy beamforming. As network scales, and need to ensure additive signal reception at the harvesting sensor, it becomes paramount to ensure this coordination through low-overhead channel state estimation, feedbacks, and cooperation among ETs.
2. ***Optimal ET Resource Management***: In DeepCharge, energy transmitters are programmable and can be allocated to a large pool such as spectrum bands, transmission powers, beam durations, and beam schedules. As a result, there are challenges to develop large-scale convex or non-convex resource allocation algorithms such as beamforming weight design for spatial distributed ETs and collaborative scheduling of energy beams, while maintaining low computational complexity and fast convergence.
3. ***Energy Transfer Hand-offs***: In DeepCharge, a hand-off might happen when a user moves from an area that covered by a group of ETs to another area that needs to be covered with a different set of ETs. The frequency of hand-off depends on the size of ETs and their associated coverage. Since the energy needs of IoT devices change based on real-time traffic, the hand-off process design needs to consider delay, jitter, and unequal priorities of the different nodes.
4. ***Beamforming Overheads***: To improve energy transfer efficiency, feedbacks, and control messages should be minimized for energy beamforming. The impact of beamforming overheads becomes more significant as IoT networks are moving towards denser deployment with more heterogeneous setups. The channel measurements from all the transmitters to all the receivers is required a priori to apply desired beamforming. The channel information must be collected at the scale of tens of milliseconds for any real-time system. With the increase in number of transmitters and receivers, the process of collecting channel information becomes a high overhead. Thus, a real-time distributed beamforming system cannot rely on explicit channel feedback and it becomes necessary to utilize advanced channel estimation techniques.
5. ***Optimal Placement of Energy Transmitters***: The ET placement problem will determine the required number of ETs and their locations while considering on the available ambient RF power and average energy needs of IoT network. Additionally, it directly impacts maximal energy coverage and capacity of the wireless charging system. A more complicated placement problem can be

considered, including multilevel clustering of ETs with hierarchical management structure.

6. ***Timely Energy Needs Monitoring and Predictions***: IoT nodes have spatially and temporally varying energy needs, based on their geographic location, participation in data forwarding, application-specific requirements and on the network topology. It is important to monitor the traffic rates and information on the energy level of each node to schedule the duration and transfer power from ET beams and that from the ambient source that is adequate to support current node operations. Furthermore, anticipating application-specific traffic demands (based in node-initiated "energy reports") can enable a safety-related energy guard and enhance the network reliability by estimation and transfer of the power in advance.
7. ***Timely Adaptability***: The most important requirement in the design of control messages such as energy updates, command queries from controller, etc., is their latency time. More specifically, while the considered network size is huge with many traffic flows, how to design messaging protocol with fast convergent rate and low latency are among the challenges that need an optimal design.
8. ***Circuit Design***: The next-generation RF energy harvesting circuits need to operate optimally on multiband of frequencies (e.g., LTE, GSM, beamforming) with respect to energy conversion efficiency (ECE). Selection of circuit components, optimization of summation boards, and enabling cognitive energy harvesting by ultra-low power switching between different frequencies are among important challenges in this area.

7 Conclusion

In this chapter, we described DeepCharge, a new architecture for next-generation wireless charging systems. Our design addresses the limitations of existing systems and provides higher levels of adaptation and reconfigurability through the use of controller, higher wireless charging rates and distances through distributed energy beamforming, and realizes higher levels of RF harvesting through design of the integrated harvester that can capture both ambient and controlled energy beams. We demonstrated our indoor and outdoor prototypes with extensive experimental measurements. In addition, we discussed the RF exposure safety limits based on FCC regulations that need to be considered in the configuration of wireless charging systems. Finally, we summarized the most important research challenges toward realizing software-defined wireless charging systems that will power the IoT revolution of tomorrow.

Acknowledgements This work is supported by the funds available through the US National Science Foundation award CNS 1452628.

References

1. Evans, D.: Cisco Whitepaper on the Internet of Things: How the Next Evolution of the Internet is Changing Everything. (2011). http://www.cisco.com/
2. IDC Report. Worldwide and regional Internet of Things (IoT) 2014–2020 forecast: A virtuous circle of proven value and demand. https://www.business.att.com
3. Ericsson Mobility Report: On the pulse of the networked society. https://www.ericsson.com
4. Lee, J.B., Chen, Z., Allen, M.G., Rohatgi, A., Arya, R.: A High Voltage Solar Cell Array As An Electrostatic MEMS Power Supply, MEMS94, IEEE Workshop Micro Electro Mechanical Systems, pp. 331–336 (1994)
5. Sangani, K.: Power solar-the sun in your pocket. Eng. Technol. **2**(8), 3638 (2007)
6. Lin, K., Yu, J., Hsu, J., Zahedi, S., Lee, D., Friedman, J., Kansal, A., Raghunathan, V., Srivastava, M.: Heliomote: enabling long-lived sensor networks through solar energy harvesting, SenSys05. In: The 3rd International Conference on Embedded Networked Sensor Systems, Nov 2005
7. Balakumar, R., Vaidehi, V., Balamuralidhar, P.: Solar energy harvesting for wireless sensor networks, CICSYN09. In: The 1st International Conference on Computational Intelligence, Communication Systems and Networks, July 2009
8. Brunelli, D., Benini, L., Moser, C., Thiele, L.: In: An Efficient Solar Energy Harvester for Wireless Sensor Nodes, DATE08. Design, Automation and Test in Europe (2008)
9. Alippi, C., Galperti, C.: In: An Adaptive System for Optimal Solar Energy Harvesting in Wireless Sensor Network Nodes, Circuits and Systems I: Regular Papers, IEEE Transactions, July 2008
10. Guilar, N., Chen, A., Kleeburg, T., Amirtharajah, R.: Integrated solar energy harvesting and storage, ISLPED06. In: The 2006 International Symposium on Low Power Electronics and Design, Oct 2006
11. Muncuk, U., Alemdar, K., Sarode, J.D., Chowdhury, K.R.: Multi-band Ambient RF energy harvesting circuit design for enabling battery-less sensors and IoTs. IEEE Internet Things J. (2018)
12. Paradiso, J.A.: Systems for human-powered mobile computing, DAC06. In: The 43rd Design Automation Conference, pp. 645–650, July 2006
13. Torah, R., Glynne-Jones, P., Tudor, M., ODonnell, T., Roy, S., Beeby, S.: Self-powered autonomous wireless sensor node using vibration energy harvesting. Meas. Sci. Technol. **19**, 8 (2008)
14. Hayakawa, M.: Electric Wristwatch with Generator, U.S. Patent, 5 001 685, Mar 1991
15. Von Buren, T., Mitcheson, P.D., Green, T.C., Yeatman, E.M., Holmes, A.S., Troster, G.: Optimization of inertial micropower generators for human walking motion, JSEN06. IEEE Sens. J. **6**(1), 2838 (2006)
16. Leonov, C.R.V., Torfs, T., Fiorini, P., Van Hoof, C.: Thermoelectric converters of human warmth for self-powered wireless sensor nodes, JSEN07. IEEE Sens. J. **7**, 650657 (2007)
17. EE Times India. http://www.eetindia.co.in
18. Leonov, V., Van Hoof, C., Vullers, R.J.M.: Thermoelectric and hybrid generators in wearable devices and clothes, BSN09. In: The 6th International Workshop on Body Sensors Networks, pp. 195–200 (2009)
19. Leonov, V., Fiorini, P.: Thermal matching of a thermoelectric energy scavenger with the ambience, ECT07. In: The 5th European Conference on Thermo-electrics, pp. 129–133, Sept 2007
20. Campana Escale, O.A.: Study of The Efficiency of Rectifying Antenna Systems for Electromagnetic Energy Harvesting, The Degree of Engineer Thesis. Escola Tecnica Superior d.Enginyeria de Telecomunicacio de Barcelona, Department de Teoria de Senyali Comunicacions, Spain, Oct 2010
21. Briles, S.D., Neagley, D.L., Coates, D.M., Freud, S.M.: Remote Down-hole Well Telemetry, U.S. Patent, No. 6766141 B1, July 2004
22. Lu, X., Wang, P., Niyato, D., Kim, D., Han, Z.: Wireless networks with RF energy harvesting: a contemporary survey. IEEE Commun. Surv. Tutor. **17**(2), 757789 (2015)

23. Lu, X., Wang, P., Niyato, D., Kim, D.I., Han, Z.: Wireless charging technologies: Fundamentals, standards, and network applications. IEEE Commun. Surv. Tutor. **18**(2), 14131452 (2016)
24. Lu, X., Niyato, D., Wang, P., Kim, D.I.: Wireless charger networking for mobile devices: fundamentals, standards, and applications. IEEE Wirel. Commun. **22**(2), 126135 (2015)
25. Harrist, D.W.: Wireless Battery Charging System Using Radio Frequency Energy Harvesting, Master of Science Thesis, University of Pittsburgh, USA (2004)
26. Tentzeris, M.M., Kawahara, Y.: Novel energy harvesting technologies for ICT applications, SAINT08. In: IEEE International Symposium on Applications and the Internet, pp. 373–376 (2008)
27. Finkenzeller, K.: RFID Handbook: Fundamentals and Applications in Contactless Smart Cards and Identification. Wiley, Chichester, Sussex, UK (2003)
28. Annala, A.L., Oy, I., Friedrich, U.: Passive Long Distance Multiple Access UHF RFID System, Palomar Project, European Commission, Public report, Project No. IST1999-10339, Nov 2002
29. Ba, H., Demirkol, I., Heinzelman, W.: Feasibility and benefits of passive RFID wake-up radios for wireless sensor networks, GLOBECOM10. In: IEEE Global Telecommunications Conference, Dec 2010
30. Ungan, T., Reindl, L.M.: Harvesting low ambient rf-sources for autonomous measurement systems, IMTC08. In: IEEE International Instrumentation and Measurement Technology Conference, Victoria, Vancouver Island, Canada, May 2008
31. Javaheri, H., Noubir, G.: iPoint: a platform-independent passive information kiosk for cell phones, SECON10. In: The 7th Annual IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks, June 2010
32. Powercast Corporation. http://www.powercastco.com/
33. Ettus Research. https://www.ettus.com/
34. Muncuk, U., Mohanti, S., Alemdar, K., Naderi, M.Y., Chowdhury, K.R.: Software-defined wireless charging of internet of things using distributed beamforming. In: ACM Conference on Embedded Networked Sensor Systems (SenSys 2016), Demo Session, Nov 2016
35. Keyrouz, S., Visser, H.J., Tijhuis, A.G.: Ambient RF energy harvesting from DTV stations. In: Loughborough Antennas and Propagation Conference (2012)
36. Parks, A.N., Smith, J.R.: Sifting through the airwaves: efficient and scalable multiband RF harvesting. In: IEEE International Conference on RFID (IEEE RFID) (2014)
37. Park, J.-Y., Han, S.-M., Itoh, T.: A rectenna design with harmonic-rejecting circular-sector antenna. IEEE Antennas Wirel. Propag. Lett. **3**(1), 52–54 (2004)
38. Scorcioni, S., Larcher, L., Bertacchini, A., Vincetti, L., Maini, M.: An integrated RF energy harvester for UHF wireless powering applications. In: IEEE Wireless Power Transfer (WPT), Perugia, vol. 2013, pp. 92–95 (2013)
39. Le, T., Mayaram, K., Fiez, T.: Efficient far-field radio frequency energy harvesting for passively powered sensor networks. IEEE J. Solid-State Circuits **43**(5), 1287–1302 (2008)
40. Nintanavongsa, P., Muncuk, U., Lewis, D.R., Chowdhury, K.R.: Design optimization and implementation for RF energy harvesting circuits. IEEE JETCAS **2**, 2433 (2012)
41. Pinuela, M., Mitcheson, P.D., Lucyszyn, S.: Ambient RF energy harvesting in urban and semi-urban environments. IEEE Trans. Microw. Theory Tech. **61**(7), 27152726 (2013)
42. London RF survey. http://www.londonrfsurvey.org
43. Liu, Z., Zhong, Z., Guo, Y.X.: Enhanced dual-band ambient RF energy harvesting with ultra-wide power range. IEEE Microw. Wirel. Compon. Lett. **25**(9), 630–632 (2015)
44. Nishimoto, H., Kawahara, Y., Asami, T.: Prototype implementation of ambient RF energy harvesting wireless sensor networks. In: Sensors, 2010 IEEE, pp. 1282–1287, Nov. 2010
45. Vyas, R.J., Cook, B.B., Kawahara, Y., Tentzeris, M.M.: E-WEHP: a batteryless embedded sensor-platform wirelessly powered from ambient digital-TV signals. IEEE Trans Microw. Theory Tech. **61**(6), 2491–2505 (2013)
46. Shegita, R., Sasaki, T., Quan, D.M., Kawahara, Y., Vyas, R.J., Tentzeris, M.M., Asami, T.: Ambient RF energy harvesting sensor device with capacitor-leakage-aware duty cycle control. IEEE Sens. J. **13** (2013)

47. Parks, A.N., Sample, A.P., Zhao, Y., Smith, J.R.: A wireless sensing platform utilizing ambient RF energy. In: 2013 IEEE Topical Conference on Power Amplifiers for Wireless and Radio Applications, Santa Clara, CA, pp. 160–162 (2013)
48. Keyrouz, S., Visser, H.J., Tijhuis, A.G.: Multi-band simultaneous radio frequency energy harvesting. In: 2013 7th European Conference on Antennas and Propagation (EuCAP), Gothenburg, pp. 3058–3061 (2013)
49. P2110B Series 850–950 MHz Power Harvester Development Kit Powercast Corp. http://www.powercastco.com/products/development-kits/
50. Masotti, D., Costanzo, A., Prete, M.D., Rizzoli, V.: Genetic-based design of a tetra-band high-efficiency radio-frequency energy harvesting system. In: IET Microwaves, Antennas & Propagation, vol. 7, no. 15, pp. 1254–1263, 10 Dec 2013
51. Pinuela, M., Mitcheson, P.D., Lucyszyn, S.: Ambient RF energy harvesting in urban and semi-urban environments. IEEE Trans. Microw. Theory Tech. **61**(7), 2715–2726 (2013)
52. Parks, A.N., Smith, J.R.: Active power summation for efficient multiband RF energy harvesting. IEEE MTT-S International Microwave Symposium, Phoenix, AZ **2015**, 1–4 (2015)
53. Stoopman, M., Keyrouz, S., Visser, H.J., Philips, K., Serdijn, W.A.: Co-Design of a CMOS rectifier and small loop antenna for highly sensitive RF energy harvesters. IEEE J. Solid-State Circuits **49**(3), 622–634 (2014)
54. Assimonis, S.D., Daskalakis, S.N., Bletsas, A.: Sensitive and efficient RF harvesting supply for batteryless backscatter sensor networks. IEEE Trans. Microw. Theory Tech. **64**(4), 1327–1338 (2016)
55. Yang, G., Ho, C.K., Guan, Y.L.: Dynamic resource allocation for multiple-antenna wireless power transfer. IEEE Trans. Signal Process. **62**(14), 35653577 (2014)
56. Sun, Q., Zhu, G., Shen, C., Li, X., Zhong, Z.: Joint beamforming design and time allocation for wireless powered communication networks. IEEE Wirel. Commun. Lett. **18**(10), 17831786 (2014)
57. Chen, X., Wang, X., Chen, X.: Energy-efficient optimization for wireless information and power transfer in large-scale mimo systems employing energy beamforming. IEEE Wirel. Commun. Lett. **2**(6), 667670 (2013)
58. Lee, S., Liu, L., Zhang, R.: Collaborative wireless energy and information transfer in interference channel. IEEE Trans. Wirel. Commun. **14**(1), 545557 (2015)
59. Haque, I.T., Abu-Ghazaleh, N.: Wireless software defined networking: a survey and taxonomy. IEEE Commun. Surv. Tutor. **18**(4), Feb 2016
60. Hu, F., Hao, Q., Bao, K.: A survey on software-defined network and OpenFlow: from concept to implementation. IEEE Commun. Surv. Tuts. **16**(4), 2181–2206 (2014)
61. Hakiria, A., Gokhale, A., Berthou, P., Schmidt, D.C., Gayraud, T.: Software-defined networking: challenges and research opportunities for future Internet. Comput. Netw. **75**, 453471 (2014)
62. Macedo, D.F., Guedes, D., Vieira, L.F.M., Vieira, M.A.M., Nogueira, M.: Programmable networks-From software-defined radio to software-defined networking. IEEE Commun. Surv. Tutor. **17**(2), 1102–1125, 2nd Quart (2015)
63. Jagadeesan, A.N., Krishnamachari, B.: Software-defined networking paradigms in wireless networks: a survey. ACM Comput. Surv. **47**(2), 111 (2014). Jan
64. Reza, M., Sivakumar, S., Nafarieh, A., Robertson, B.: A comparison of software defined network (SDN) implementation strategies. In: Proceedings of the 2nd International Workshop Survivable Robust Optical Network, Hasselt, Belgium, pp. 1050–1055, Jun 2014
65. Kreutz, D., Ramos, F.M.V., Verssimo, P.E., Rothenberg, C.E., Azodolmolky, S., Uhlig, S.: Software-defined networking: a comprehensive survey. In: Proceedings of the IEEE, vol. 103, no. 1, pp. 14–76
66. Akyildiz, I.F., Wang, P., Lin, S.-CH.: SoftAir: a software defined networking architecture for 5G wireless systems. Comput. Netw. J. **85**, 1–18, July 2015
67. Jain, S., Kumar, A., Alok, M., Mandal, S., Ong, J., Poutievski, L., Leon, S., Arjun, V., Venkata, S., Wanderer, J., Jim, Z., Zhou, J., Zhu, M., Zolla, J., Holzle, U., Stuart, S., Vahdat, A.: B4: experience with a globally-deployed software defined wan. SIGCOMM Comput. Commun. Rev. **43**(4) (2013)

68. Akyildiz, I.F., Wang, P., Lin, S.-C.H.: SoftWater: Software-defined networking for next-generation underwater communication systems. Ad Hoc Netw. J. **46**, 111 (2016)
69. Cao, B., He, F., Li, Y., Wang, C., Lang, W.: Software defined virtual wireless network: framework and challenges. IEEE Netw. **29**(4), 612, Jul/Aug 2015
70. Chen, L., Warner, J., Yung, P.L., Zhou, D., Heinzelman, W., Dermirkol, I., Muncuk, U., Chowdhury, K.R., Basagni, S.: REACH2-Mote: a range extending passive wake-up wireless sensor node. ACM Trans. Sens. Netw. **11**(4) (2015)
71. Kaushik, K., Mishra, D., De, S., Chowdhury, K.R., Heinzelman, W..: Low-Cost Wake-Up receiver for RF energy harvesting wireless sensor networks. IEEE Sens. J. **16**(16) (2016)
72. Cid-Fuentes, R.G., Naderi, M.Y., Basagni, S., Chowdhury, K., Cabellos-Aparicio, A., Alarcon, E.: On Signaling Power: Communications over Wireless Energy. IEEE INFOCOM, San Francisco, CA, USA (2016)
73. Cid-Fuentes, R.G., Naderi, M.Y., Basagni, S., Chowdhury, K.R., Cabellos-Aparicio, A., Alarcon, E.: An All-Digital Receiver for Low Power, Low Bit-Rate Applications Using Simultaneous Wireless Information and Power Transmission, IEEE ISCAS 2016, Montreal, Canada, May 2016
74. Naderi, M.Y., Chowdhury, K.R., Basagni, S., Heinzelman, W., De, S., Jana, S.: Surviving wireless energy interference in RF-harvesting sensor networks: an empirical study. In: IEEE SECON Workshop on Energy Harvesting Communications, Singapore (2014)
75. Kaushik, K., Mishra, D., De, S., Basagni, S., Heinzelman, W., Chowdhury, K.R., Jana, S.: Experimental Demonstration of Multi-Hop RF Energy Transfer. IEEE PIMRC, London, UK (2013)
76. Doost, R., Chowdhury, K.R., DiFelice, M.: Routing and link layer protocol design for sensor networks with wireless energy transfer. In: Proceedings of IEEE Globecom, Miami, Fl (2010)
77. Chen, L., Cool, S., Ba, H., Heinzelman, W., Demirkol, I., Muncuk, U., Chowdhury, K.R., Basagni, S.: Range extension of passive wake-up radio systems through energy harvesting. In: Proceedings of IEEE ICC, Budapest, Hungary, June 2013
78. Naderi, M.Y., Chowdhury, K.R., Basagni, S.: Wireless sensor networks with RF energy harvesting: energy models and analysis. In: IEEE WCNC, Accepted, New Orleans, LA (2015)
79. Mishra, D., Kaushik, K., De, S., Basagni, S., Chowdhury, K.R., Jana, S., Heinzelman, W.: Implementation of multi-path energy routing. In: IEEE PIMRC, Washington DC, Sept 2014
80. Cid-Fuentes, R.G., Naderi, M.Y., Doost, R., Chowdhury, K.R., Cabellos-Aparicio, A., Alarcon, E.: Leveraging deliberately generated interferences for multi-sensor wireless RF power transmission. In: Proceedings of IEEE GLOBECOM, San Diego, CA, USA, p. 2015, Dec 2015
81. Naderi, M.Y., Basagni, S., Chowdhury, K.R.: Modeling the residual energy and lifetime of energy harvesting sensor nodes. In: Proceedigns of IEEE GLOBECOM, Anaheim, CA, USA, Dec 2012
82. De, S., Mishra, D., Chowdhury, K.R.: Charging time characterization for wireless RF energy transfer. IEEE Trans. Circuits Syst. II **64**(4) (2015)
83. Mishra, D., De, S., Jana, S., Basagni, S., Chowdhury, K.R., Heinzelman, W.: Smart RF energy harvesting communications: challenges and opportunities. IEEE Commun. Mag, Accept (2014)
84. Naderi, M.Y., Chowdhury, K.R., Basagni, S., Heinzelman, W., De, S., Jana, S.: Experimental study of concurrent data and wireless energy transfer for sensor networks. In: IEEE GLOBECOM, Austin, TX (2014)
85. Nintanavongsa, P., Naderi, M.Y., Chowdhury, K.R.: A dual-band wireless energy transfer protocol for heterogeneous sensor networks powered by RF energy harvesting. In: IEEE International Computer Science and Engineering Conference (ISCEC), Bangkok, Thailand, Sept 2013
86. Naderi, M.Y., Nintanavongsa, P., Chowdhury, K.R.: RF-MAC: a medium access control protocol for re-chargeable sensor networks powered by wireless energy harvesting. IEEE Trans. Wirel. Commun. **13**(7), July 2014
87. Coarasa, A.H., Nintanavongsa, P., Sanyal, S., Chowdhury, K.R.: Impact of mobile transmitter sources on radio frequency wireless energy harvesting. In: Proceedings of IEEE International Conference on Computing, Networking and Communications (ICNC), San Diego, CA, Jan 2013

88. FCC Radio Frequency Safety Guidelines. https://www.fcc.gov/general/radio-frequency-safety-0
89. HPA-850 RF Bay Amplifier. http://rfbayinc.com/
90. FCC RF Exposure Wireless Charging Apps v02. https://apps.fcc.gov/eas/comments/GetPublishedDocument.html?id=319&tn=270151
91. FCC General RF Exposure Guidance v06, FCC publication number: 447498. https://apps.fcc.gov/oetcf/kdb/forms/FTSSearchResultPage.cfm?switch=P&id=20676
92. Mudumbai, R., Hespanha, U.M.J., Barriac, G.: Distributed transmit beamforming using feedback control. IEEE Trans. Inf. Theory 411426 (2010)
93. Mudumbai, R., Brown, D.R., Madhow, U., Poor, H.V.: Distributed transmit beamforming: challenges and recent progress. IEEE Commun. Mag. **47**(2), 102110 (2009)
94. Yan, H., Macias Montero, J.G., Akhnoukh, A., de Vreede, L.C.N., Burghart, J.N.: An integration scheme for RF power harvesting. In: The 8th Annual Workshop on Semiconductor Advances for Future Electronics and Sensors, Veldhoven, Netherlands (2005)

Part VI
Robotics and Middleware

Robotic Wireless Sensor Networks

Pradipta Ghosh, Andrea Gasparri, Jiong Jin and Bhaskar Krishnamachari

Abstract In this chapter, we present a literature survey of an emerging, cutting-edge, and multidisciplinary field of research at the intersection of Robotics and Wireless Sensor Networks (WSN) which we refer to as *Robotic Wireless Sensor Networks (RWSN)*. We define an *RWSN* as an autonomous networked multi-robot system that aims to achieve certain *sensing goals* while meeting and maintaining certain *communication performance requirements*, through cooperative control, learning, and adaptation. While both of the component areas, i.e., robotics and WSN, are very well known and well explored, there exist a whole set of new opportunities and research directions at the intersection of these two fields, which are relatively or even completely unexplored. One such example would be the use of a set of robotic routers to set up a temporary communication path between a sender and a receiver that uses the controlled mobility to the advantage of packet routing. We find that there exist only a limited number of articles to be directly categorized as RWSN-related works whereas there exist a range of articles in the robotics and the WSN literature that are also relevant to this new field of research. To connect the dots, we first identify the core problems and research trends related to RWSN such as connectivity, localization, routing, and robust flow of information. Next, we classify the existing research on RWSN as well as the relevant state of the arts from robotics and WSN community according to the problems and trends identified in the first step. Lastly, we analyze what is missing in the existing literature and identify topics that require more research attention in the future.

P. Ghosh (✉) · B. Krishnamachari
University of Southern California, Los Angeles, CA 90089, USA
e-mail: pradiptg@usc.edu

B. Krishnamachari
e-mail: bkrishna@usc.edu

A. Gasparri
Universit degli studi "Roma Tre", Via della Vasca Navale, 79, 00146 Roma, Italy
e-mail: gasparri@dia.uniroma3.it

J. Jin
Swinburne University of Technology, Melbourne, VIC 3122, Australia
e-mail: jiongjin@swin.edu.au

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_16

1 Introduction

Robotics has been a very important and active field of research over last couple of decades with the main focus on seamless integration of robots in human lives to assist and to help human in difficult, cumbersome jobs such as search and rescue in disastrous environments and exploration of unknown environments [1, 2]. The rapid technological advancements over last two decades in terms of cheap and scalable hardware with necessary software stacks have provided a huge momentum to this field of research. As part of this increasing stream of investigations into robotics, researchers have been motivated to look into the collaborative aspects where a group of robots can work in synergy to perform a set of diverse tasks [3, 4]. Nonetheless, most of the research works on collaborative robotics, such as swarming, have remained mostly either theoretical concepts or incomplete practical systems which lack some very important pieces of the puzzle such as realistic communication channel modeling and efficient network protocols for interaction among the robots. Note that, we use the term "realistic communication channel model" to refer to a wireless channel model that accounts for most of the well-known dynamics of a standard wireless channel such as path loss, fading, and shadowing [5]. On the other hand, the field of wireless networks (more specifically, Wireless Sensor Networks (WSN) and wireless ad hoc networks) has been explored extensively by communication and network researchers where the nodes are considered static (sensor nodes) or mobile without control (mobile ad hoc network). With the availability of cheap easily programmable robots, researchers have started to explore the advantages and opportunities granted by the controlled mobility in the context of wireless networks. Nonetheless, the mobility models used by the network researchers remained simple and impractical, and not very pertinent to robotic motion control until last decade.

Over last decade, a handful of researchers noticed the significant disconnection between the robotics and the wireless network research communities and its bottleneck effects in the full-fledged development of a network of collaborative robots. Consequently, researchers have tried to incorporate wireless network technologies in robotics and vice verse, which opened up a whole new field of research at the intersection of robotics and wireless networks. This new research domain is called by many different names such as "*Wireless Robotics Networks*", "*Wireless Automated Networks*", and "*Networked Robots*". In this chapter, keeping in mind that the primary task of teams of robots in many application contexts might be pure sensing, we will refer to this field as "*Robotic Wireless Sensor Networks (RWSN)*". According to the IEEE Society of Robotics and Automation's Technical Committee: "*A 'networked robot' is a robotic device connected to a communications network such as the Internet or LAN. The network could be wired or wireless, and based on any of a variety of protocols such as TCP, UDP, or 802.11. Many new applications are now being developed ranging from automation to exploration. There are two subclasses of Networked Robots: (1) Tele-operated, where human supervisors send commands and receive feedback via the network. Such systems support research, education, and public awareness by making valuable resources accessible to broad audiences; (2)*

Autonomous, where robots and sensors exchange data via the network. In such systems, the sensor network extends the effective sensing range of the robots, allowing them to communicate with each other over long distances to coordinate their activity. The robots in turn can deploy, repair, and maintain the sensor network to increase its longevity, and utility. A broad challenge is to develop a science base that couples communication to control to enable such new capabilities". We define an *RWSN* as an autonomous networked multi-robot system that aims to achieve certain *sensing goals* while meeting and maintaining certain *communication performance requirements* via cooperative control, learning, and adaptation. Another important definition related to this field is "*cooperative behavior*" which is defined as follows: "*given some task specified by a designer, a multiple-robot system displays cooperative behavior if, due to some underlying mechanism (i.e., the 'mechanism of cooperation'), there is an increase in the total utility of the system*." A group of cooperative robots is of more interest than single robot because of some fundamental practical reasons such as easier completion and performance benefits of using multiple simple, cheap, and flexible robots for complex tasks.

Over the years, robotics and wireless network researchers have developed algorithms as well as hardware solutions that directly or indirectly fall under the umbrella of RWSN. Network of robots is already experimentally applied and tested in a range of applications such as Urban Search And Rescue missions (USAR), firefighting, underground mining, and exploration of unknown environments. The very first practical USAR mission was launched during the rescue operations at the World Trade Center on September 11, 2001 [2] using a team of four robots. In the context of firefighting, a group of Unmanned Aerial Vehicles (UAV) was used for assistance at the Gestosa (Portugal) forest fire in May 2003 by Ollero et al. [6]. Communication links between the robots in such contexts can be very dynamic and unreliable, thereby, require special attention for an efficient operation. This requires careful movement control by maintaining good link qualities among the robots. Among other application contexts, underground mining is very important. Due to many difficulties like lack of accurate maps, lack of structural soundness, harshness of the environment (e.g., low oxygen level), and the danger of explosion of methane, accidents are almost inevitable in underground mining resulting in the deaths of many mine workers. Thrun et al. [7] developed a robotic system that can autonomously explore and acquire three-dimensional maps of abandoned mines. Later, Murphy et al. [8] and Weiss et al. [9] also presented models and techniques for using a group of robots in underground mining. The field of cooperative autonomous driving, one of the major research focuses in the automobile industry, also falls under the broad umbrella of RWSN. Baber et al. [10], Nagel et al. [11], Milanes et al. [12], and Xiong et al. [13] worked on solving a range of problems in practical implementations of autonomous driving systems and flocking of multiple unmanned vehicles like Personal Air Vehicle (PAV). Robot swarms [14, 15], which deal with large numbers of autonomous and homogeneous robots, are also special cases of RWSN. Swarms have limited memory and have very limited self-control capabilities. Example use cases of swarms are in searching and collecting tasks (food harvesting [16], in collecting rock samples on distant planets [17]), or in collective transport of palletized loads [18]. Penders et al.

[1] developed a robot swarm to support human in search missions by surrounding them and continuously sensing and scanning the surroundings to inform them about potential hazards. A swarm of robots can also be used in future health-care systems, e.g., swarm of microrobots can be used to identify and destroy tumor/cancer cells. Military application is another obvious field of application. Many researchers have been working on developing military teams of autonomous UAVs, tanks and Robots, e.g., use of Unmanned Ground Vehicles (UGV) during RSTA missions. One example of such project is "Mobile Autonomous Robot Systems (MARS)," sponsored by DARPA. The works of Nguyen et al. [19] and Hsieh et al. [20] are mentionable on military application of Networked Robots. In the field of Exploration, the most famous example is the twin Mars Exploration Rover (MER) vehicles. They landed on Mars in the course of January 2004 [21]. Among other applications, hazardous waste management, robot sports [22], mobile health-care [23], smart home [24–26], smart antenna, deployment of communication network and improvement of current communication infrastructure [27, 28], and cloud networked robotics [29–31] are also important. In summary, there exists a huge range of applications of an RWSN. In Table 1, we list the different types of applications of an RWSN.

In this chapter, we present a literature survey of the existing state-of-the-arts on RWSN. We discover that while there exist significant amount of works [32] in both the ancestral fields (robotics and WSN) that are also relevant to RWSN, most of these state of the arts cannot be directly classified as RWSN-related works, yet should not be omitted. Nonetheless, there also exist a range of works that properly lie at the intersection of robotics and WSN and, therefore, directly fall under the purview of RWSN. To draw a complete picture of the RWSN-related state of the arts, we **first** identify and point out current research problems, trends, and challenges in the field of RWSN such as connectivity, localization, routing, and robust flow of information. While some of these problems are inherited from the fields of robotics (such as path control and coordination) and WSN (such as routing and localization), a new class of independent problems have also emerged such as link quality maintenance, Radio

Table 1 Application summary of RWSN

Applications	References
Search and rescue	[1, 2, 6, 10]
Mining	[7–9]
Autonomous driving	[10–13]
Robot swarm	[14–18]
Military applications	[19–21]
Robot sports	[22]
Mobile health-care	[23]
Smart home	[24–26]
Deployment of communication network	[27, 28]
Cloud networked robotics	[29–31]

Signal Strength Information (RSSI) estimations in present and future location of a robot, and guaranteed proximity between neighboring nodes. **Second**, we categorize the existing research in accordance with the problems they address. In doing so, we also include solutions that are not directly applicable but provide with a solution base to build upon. For example, in the contexts of connectivity maintenance, the traditional solutions involve representing the network as a graph by employing simple unit disk wireless connectivity model [33]. However, in practical employment, the wireless communication links do not follow unit disk model and rather follow a randomized fading and shadowing model [5]. Thus, such existing solutions do not directly fall under RWSN yet can be modified to include more realistic communication model and, thus, should not be omitted. **Lastly**, we discuss what is missing in the existing literature, if any, as well as some potential directions of future research in the field of RWSN. *In this chapter, we further discuss how and where do the existing works fit in the context of the layered architecture (Internet model) of a network stack in order to identify key networking goals and problems in an RWSN.*

2 What is an RWSN?

In this section, we illustrate the field of RWSN in detail. Here, we address some core and important questions related to RWSN: What is an RWSN? What kind of research works are classified as RWSN-related works?

We define an RWSN as a wireless network that includes a set of robotic nodes with controlled mobility and a set of nodes equipped with sensors, whereas all nodes have wireless communication capabilities. Ideally, each node of a robotic sensor network should have controlled mobility, a set of sensors, and wireless communication capabilities (as illustrated in Fig. 1). We refer to such nodes (devices) as "**Robotic Wireless Sensors (RWS)**". Nonetheless, an RWSN can also have some nodes with just sensing and wireless communication capabilities but without controlled mobilities. We refer to such nodes (by following traditional terminology) as "**Wireless Sensors**". Note that, every node of an RWSN must have wireless communication capabilities according to our definition. Moreover, an RWSN is typically expected to be able to fulfill or guarantee certain communication performance requirements enforced by the application contexts such as minimum achievable Bit Error Rate (BER) in every link of the network.

To answer the second question, the existing research works related to RWSN can be subdivided into two broader genres. The first genre focuses on generic multi-robot sensing systems with realistic communication channels (i.e., including the effects of fading, shadowing, etc.) between the robots. *To clarify, these are mostly the existing works in the robotics literature on multi-robot systems but with practical wireless communication and networking models*. One application context of such an RWSN is in robot-assisted firefighting where the robots are tasked to sense the unknown environments inside rubble to help and guide the firefighters. Now, if the robots are not able to maintain a good connectivity among themselves or to a mission

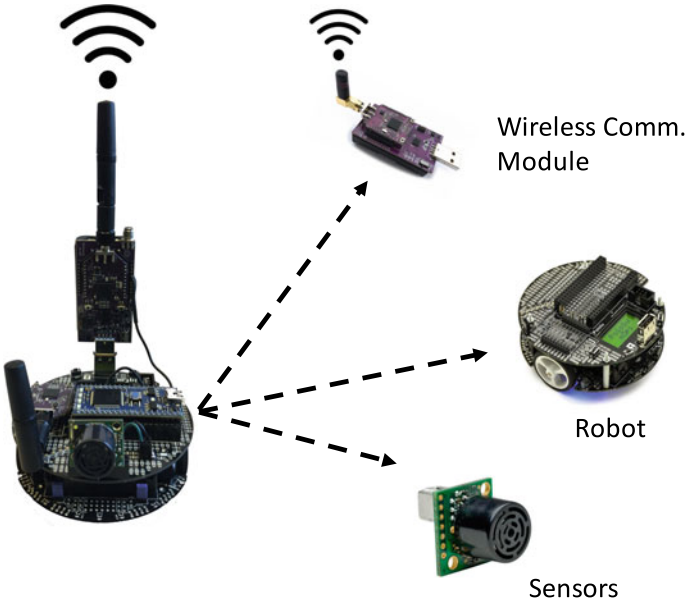


Fig. 1 Illustration of a robotic wireless sensor

control station, the whole mission is voided. Refer to Fig. 2 for an illustration of such contexts where a group of robots is sensing an unknown environment to guide the human movements. In Fig. 2, the network consisting of five robots and two firefighters needs to be connected all the time and also needs to have properties such as reliability and lower packet delays. Thus, we need a class of multi-objective motion control that will optimize the sensing and exploration task performance, and will also ensure the connectivity and the performance of the network. Some of the main identifiable challenges in this genre of works are as follows: connectivity maintenance, efficient routing to reduce end-to-end delay of packets, and multi-objective motion control and optimization.

The second genre of works focuses on the application of RWS to create and support a temporary communication backbone between a set of communicating entities. In these contexts, we sometimes use the terms "**robotic router**" and "**robotic wireless sensor**" synonymously, to put emphasis on the communication and routing goals. The main theme of these works is to exploit the controlled mobility of the robotic routers to perform sensing and communication tasks. *Note that, there exists a vast literature on multi-agent systems in robotics and control community that apply simple disk models for communication modeling and, subsequently, apply graph theory to solve different known problems such as connectivity and relay/repeater node placements. In order to be directly included in the RWSN literature, these existing works need to include the effects of fading and shadowing in the communication models which is likely to significantly increase the complexity of the problems as well as the*

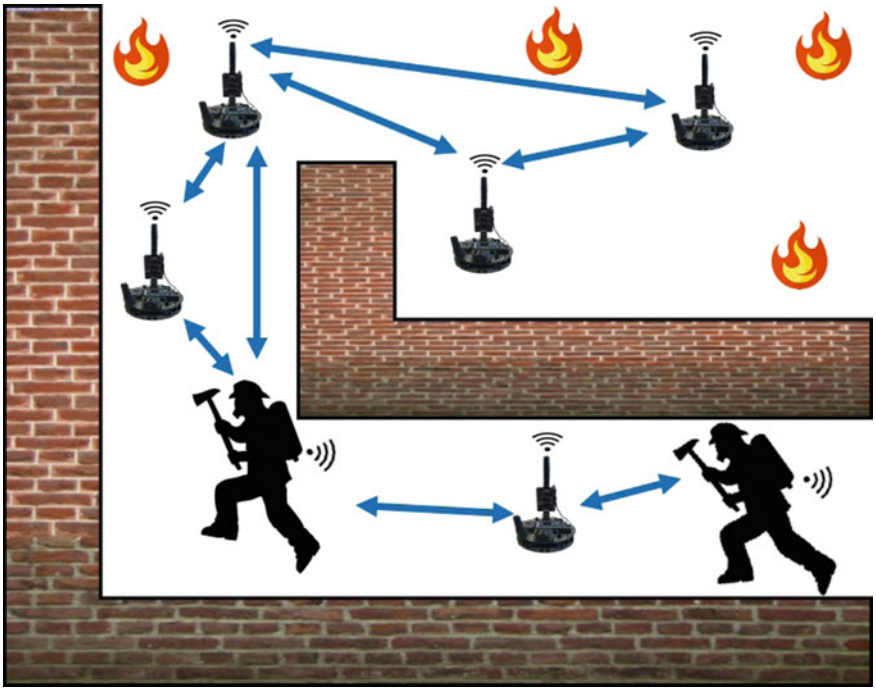


Fig. 2 Illustration of robotic sensor where a group of five robots is sensing the environment around the firefighters to guide them in firefighting while also providing connectivity

solutions. An example of the second genre of works is in the application of RWS in setting up a temporary communication backbone. While sensing is still involved for the robotic router placement optimization and adaptation, the main purpose of the system is to support communication, not sensing. In Fig. 3, we present an example illustration where a set of two robotic routers form a communication relay path between two humans (e.g., two firefighters) who are unable to communicate directly. Some of the main challenges in this genre of RWSN research are as follows: link performance guarantee (in terms of Signal to Interference plus Noise Ratio, SINR, or Bit Error Rate, BER), optimized robotic router placements and movements in a dynamic network, nonlinear control dynamics due to inclusion of network performance metrics into control loop, and localization. A special case of this would be robot-assisted static relay deployments, where the robots act as carriers of static relay nodes and smartly place/deploy them to form a communication path/backbone.

Next, we identify a set of system components and algorithms required in an RWSN as follows. All these pieces are individual research problems themselves and thus require separate attention.

- **RSSI Models, Measurements, and RF Mapping**: In an RWSN, it is important to estimate and monitor the quality of the communication links between the nodes

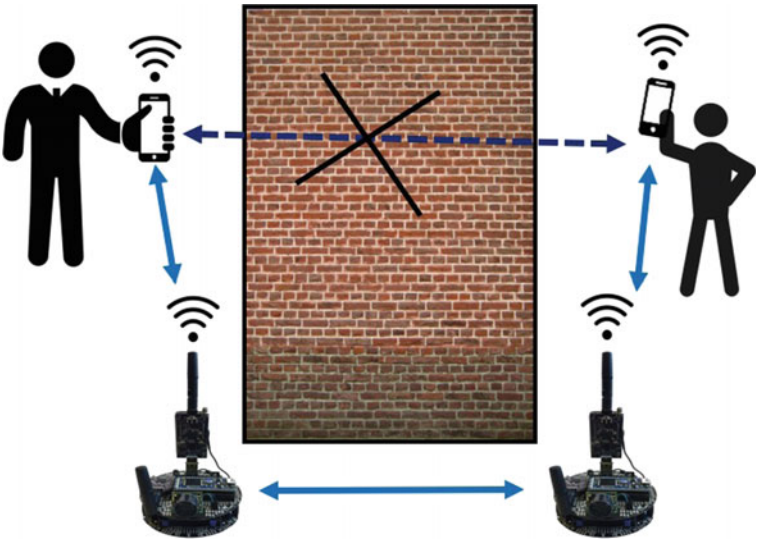


Fig. 3 Illustration of robotic routers where the two humans are not able to communicate directly due to the presence of a wall or some other blocking objects

(in terms of Bit Error Rate (BER), Signal to Noise plus Interference Ratio (SINR), etc.) in order to satisfy the communication-related requirements. (Note that, RF-based communication is standard mode of communication in RWSN for obvious reasons.) For practicality, these estimations must be either partly or fully based on online RF sensing such as temporal RSSI measurements in a deployment. Moreover, in some application contexts of RWSN, the sole goal of a robotic sensor network can be to sense and formulate an RF map of an environment to be processed or exploited later on. We present a survey of such RF mapping and modeling-related works in Sect. 3.1.

- **Routing Protocols**: Similar to any wireless sensor networks, routing and data collection is crucial in an RWSN. The concept of RWSN has opened up the door to a new class of routing protocols that incorporates the controlled mobility of the nodes in the routing decisions for more effective communication. Moreover, end-to-end delay reduction and reliability improvement have become of prime interests. A brief survey of existing RWSN-related works on routing protocols is presented in Sect. 3.2.
- **Connectivity Maintenance**: In any collaborative network of robots, it is important to maintain a steady communication path (direct or multi-hop) between any pair of nodes in the network for an effective operation. This problem, traditionally referred to as connectivity maintenance problem, in very well studied by the robotics research community. A survey of such state of the arts is presented in Sect. 3.3.

- **Communication-Aware Robot Positioning and Movement Control**: As mention earlier, one of the application contexts of RWSN is in supporting temporary communication backbones. The most important research question in such contexts is to devise a control system that adapts the positions of the robotic routers throughout the period of deployment to optimize the network performance while optimizing the movements as well. Therefore, the main goal of these class of work is continuous joint optimization of the robotic movements and the wireless network's performance. Moreover, the router placement controller should also be able to support network dynamics such as node failures and change in the set of communication endpoints. Another important application context of RWSN systems is distributed coordinated sensing using multiple robots. In such sensing contexts, the robotic sensing agents should be able to optimally sense the region of interest and route the sensed data to other nodes or a command center. This also requires careful communication-channel-dynamics-aware positioning of the robotic sensors. We present a summary of such communication-aware robotic router/sensor positioning works in Sect. 3.4.
- **Localization**: Localization is a well-known problem in the field of WSN as well as robotics. Thus, it is quite intuitive to be included in the field of RWSN. Moreover, the field of RWSN sometimes requires techniques for robots to follow each other or maintain proximity to each other. For that high accuracy relative localization is more important than absolute localization. We present a brief summary of such localization-related works in Sect. 3.5.

3 RWSN System Components

In this section, we present a categorical survey of all state of the arts in the field of RWSN. The works are classified according to the problem addressed.

3.1 RSS Models, Measurements, and RF Mapping

Radio Signal Strength (RSS) variation over a spacial domain greatly impacts the wireless communication properties, such as power decay and packet loss, between two nodes. Different properties of a physical environment, such as obstacles and propagation medium, affect radio signal propagation in different way, thereby, causing fading, shadowing, interference, and path loss effects [5]. All these should be taken into consideration (via proper communication channel models) for proper selection of radio transmission and reception parameters to improve the communication quality in an RWSN. While there exists a range of standard communication channel models in the literature, such as simple path loss model and log-normal fading model [5], the

applicability as well the model parameters' values depend on the actual deployment environments. For example, according to a log-normal fading model [5], the received power is calculated as follows:

$$\begin{aligned} P_{r,dBm} &= P_{t,dBm} + G_{dB} - L_{ref} - 10\eta \log_{10} \frac{d(t)}{d_{ref}} + \psi \\ P^{ref}_{r,dBm} &= P_{t,dBm} + G_{dB} - L_{ref} + \psi \end{aligned} \tag{1}$$

where $P_{r,dBm}$ is the received power in dBm, $P_{t,dBm}$ is the transmitter power in dBm, G_{dB} is the gain in dB, L_{ref} is the loss at the reference distance d_{ref}, η is the path loss exponent, $d(t)$ is the distance between the transmitter and receiver, $\psi \sim \mathcal{N}(0, \sigma^2)$ is the random shadowing and multipath fading noise which is log normal with variance σ^2, and $P^{ref}_{r,dBm}$ is the received power at reference distance (d_{ref}). While some of the variables such as $P_{t,dBm}$, d_{ref}, and G_{dB} are known or can be measured, the values of other variables such as η, L_{ref}, and σ are dependent upon the deployment environment. Thus, it is important to identify or estimate the proper values of such parameters in the deployment environment in order to estimate the $P_{r,dBm}$. Moreover, RSS models and maps are very important and useful database that can be used for a range of purposes such as localization of nodes based on received signal strength [34–37], mapping of an unknown terrain [38], and identifying obstacles [39]. The communication model-based estimations of the RF signal strengths in the future and unvisited locations are also very important for maintaining connectivity among mobile nodes and for optimizing the network performance. In this section, we provide a brief overview of the state-of-the-art RSS modeling and mapping techniques that are applicable in an RWSN.

What Is Already Out There? In the context of mapping and modeling of the RF channels, the most common and practical class of approaches is to deploy the network of robots with an initial model of the communication channel. Then, the robots continually or periodically collect RF samples to update the communication model parameters in an online fashion. One can also opt for an offline modeling where the robots collect a set of RF samples over the region of interest followed by post-processing of all the samples to estimate the communication channel properties. Nonetheless, the later class of methods is unrealistic and of little interest to us as it can not cope with temporal changes in the communication channel properties, which is a quite common phenomenon.

One of the key challenges of online RF mapping is in the sparsity of RF samples. Mostofi et al. [40] have done significant research in exploiting this sparsity to their advantage for RF channel modeling in networked robot systems. In [41], Mostofi and Sen presented a technique of RSS mapping by exploiting sparse representation of the communication channel in frequency domain. They demonstrated how one can reconstruct the original signal using only a small number of sensing measurements by employing the theory of *compressive sampling* [42]. Later, they utilized the compressive sampling-based reconstruction of the signal in the domain of cooperative mapping [38] to build a map of the region of interest. Mostofi et al.

[43] further presented an overview of the characterization and modeling of wireless channels for an RWSN. In their works, all three major dynamics in a physical wireless communication channel (small-scale fading, large-scale fading, and path loss) are considered [44–46]. In [46], Mostofi, Malmirchegini, and Ghaffarkhah also presented techniques for channel predictions in future locations of robots based on the compressed sensing-based channel models. Since all these are completely based on wireless measurements, multipath fading has a great influence over the results. Later on, Yan and Mostofi [47] presented a combined framework for the optimization for motion planning and communication planning. The concept of compressive sensing-based signal reconstruction and material-dependent RF propagation properties are also employed for RF-based imaging and mapping of cluttered objects/regions [39, 48, 49]. In these works, RF signal propagation properties (mostly attenuations) of the communication path between pairs of moving robots are used to estimate a structural map of an obstacle or a cluttered region. Hsieh et al. [50] have also demonstrated an RSS mapping technique using a group of robots in an urban environment. The goal was to learn the environment's communication characteristics to generate a connectivity graph. Later on, they used this model for connectivity maintenance problem in a network of robots [51]. In [52], Fink and Kumar presented another mapping technique using multiple robots. Their goal was to use the mapping for localization of an unknown source using a Gaussian process-based maximum likelihood detection. They further extended the Gaussian process-based channel model to guarantee a minimum stochastic end-to-end data rate in a network of robots [53]. Signal attenuation factors due to the presence of obstacles are taken into account in the work of Wang, Krishnamachari, and Ayanian [54]. Ghosh and Krishnamachari [55] have proposed a log-normal model-based stochastic bound on interference power and SINR for any RWSN. The works presented in [56–58] are also related to this context. A concise summary of all these related work is presented in Table 2.

What Would Be the Potential Future Research Directions? To our knowledge, a generic model for interference and SINR estimation is missing in the current literature. Interference modeling is a key to develop a more realistic model of wireless channels. Interference should also be taken into account for robotic router placements contexts, where the main goal of the network is to guarantee certain communication performance qualities. In our opinion, this should be a major focus of future research in this topic. Moreover, the effects of channel access protocols, such as Carrier Sense Multiple Access (CSMA) [5], need to be taken into account in the interference and SINR models. The existing signal strength models could be extended to achieve this goal. Nonetheless, to our knowledge, this has remained an unexplored problem in the context of an RWSN. Another future research direction would be to perform an extensive set of measurement experiments in real (rather than simulated) mines, undergrounds, or firefighting environments. This data can be used to identify the RF properties in such environments that can be exploited to the advantage of RWSN system design.

Table 2 Summary of RSS models, measurements, and RF mappings-related works

Algorithm classes	**Online**: Modeling during deployment [38, 40, 41, 43, 46, 47]
	Offline: Modeling before deployment [50, 51, 55]
Challenges	Sparse sampling [38, 40]
	Future location's signal prediction [46, 47]
	Temporal dynamics [46]
Available theoretical tools	Compressive sampling [42]
	Gaussian process [52, 55]
	Fading models [5]
	Path loss models [5]
Potential future directions	Interference and SINR models
	Account for channel access methods like CSMA
	Real-world data collection and analysis

3.2 Routing

Routing in an RWSN can be considered same as in Mobile Ad Hoc Networks (MANET) but with an extra advantage of controllability. In MANET, there are mainly two types of popular routing algorithms called reactive and proactive techniques. Among the reactive techniques, Ad Hoc On-Demand Distance Vector (AODV) [59], [60] and Dynamic Source Routing (DSR) [61], [62] are the most popular ones. On the other hand, in the class of proactive techniques, Optimized Link State Routing (OLSR) [63] and B.A.T.M.A.N. [64] are the popular ones. While any of these algorithms can be used for an RWSN, they do not take advantage of the extra feature in RWSN: controlled mobility. Ideally, the controlled mobility aspect should also be taken into account for optimized routing decisions. Nonetheless, it remained to be one of the less explored areas in RWSN. Moreover, a lower end-to-end delay and higher reliability in packet routing (mostly control packets) are two important and required aspects in an RWSN. Delayed or missing packets can result in an improper collaborative movement control and task completion in an RWSN. To this extent, some researchers have modified existing routing solutions to adapt in a robotic network and proposed completely new routing solutions as well. In this section, we present a brief survey of the state-of-art routing techniques in RWSN that are developed or modified with sole focus on robotics.

MRSR, MRDV, and MRMM: In [65], Das et al. presented three routing protocols based on traditional mobile ad hoc protocols, such as DSR [61] and AODV [60], for routing in a network of mobile robots. A brief description of each of these algorithms is as follows:

- *Mobile Robot Source Routing (MRSR)*: It is a unicast routing algorithm based on Dynamic Source Routing (DSR) [66]. MRSR incorporates three mechanisms: *route discovery*, *route construction*, and *route maintenance*. In the runtime of *route discovery* phase, each robot along the pathway of *route reply* message encodes its mobility information into the route reply packet. During *route construction*, MRSR exploits graph cache that contains the topological information of the network. The *route maintenance* phase is similar to the maintenance method applied in DSR.
- *Mobile Robot Distance Vector (MRDV)*: This is also a unicast routing algorithm based on the well-known AODV [67] routing protocol. MRDV protocol adopts AODV features such as the on-demand behavior and hop-by-hop destination sequence number. Nevertheless, unlike MRSR, MRDV explores only one route that may not have the longest lifetime among all possible routes, thereby, resulting in high probability of route errors.
- *Mobile Robot Mesh Multicast (MRMM)*: This is a multicast protocol for mobile robot networks based on ODMRP (On-Demand Multicast Routing Protocol) [68] for MANETs.

Adaptive Energy Efficient Routing Protocol (AER): This protocol was proposed by Abishek et al. [69] to achieve optimal control strategy for performing surveillance using a network of flying robots. AER protocol is also subdivided into three phases (similar to DSR [66]): *route discovery*, *route maintenance*, and *route failure handling*. The residual energy levels and signal strengths at the neighboring nodes are the main route determining factors in this protocol. To model them, the authors defined two decision parameters: T (attribute value of the neighbor) and C (cost function). The best value of T decides the forwarding node. The nodes that are neither selected for message forwarding nor have sufficient energy, are switched to the sleep state to minimize energy consumptions.

ACTor-based Robots and Equipment Synthetic System (ACTRESS): In [70], Matsumoto et al. proposed a robotic platform called ACTRESS that consists of robotic elements referred to as *robotors*. They also proposed a routing protocol exclusively for that platform. The messages are classified into two different classes: messages to establish/relinquish a communication link, and messages for control and rest of the purposes. The first kind of messages uses traditional communication protocols such as TCP/IP. The second type of messages uses its proposed special protocol to establish logical links, allocate tasks, and control cooperative motions. For this purpose, the authors have introduced four levels of messages: physical level, procedural level, knowledge level, and concept level. The common part of all four types of messages is referred to as the *message protocol core*, which is used for: negotiation, inquiry, offer, announcement and synchronization.

WNet: Tiderko et al. [71] proposed a new multicast communication technique called WNet that is based on the well-known Optimized Link State Routing(OSLR) protocol [63]. Similar to OLSR, WNet uses HELLO and Topology Control (TC) management frames to create and update the network topology graph stored in each robot node. However, the packet frames are integrated with some additional infor-

mation of link attributes that is used for link quality estimation. Next, the Dijkstra algorithm [72] is applied to the topology graphs to determine the routing paths.

Steward-Assisted Routing (StAR): In [73], Weitzenfeld et al. presented a new routing algorithm called StAR that deals with mobility and interference in an RWSN. The objective of this protocol is to nominate, for each connected partition, a "*steward*" for each destination. These *stewards* are noting but next hop robots toward destination that can store the data until a route to the destination is available. The message routing of StAR is based on a combination of global contact information and local route maintenance. Also, periodic broadcast messages with unique source identifiers are sent containing topological location of the active destination. At the beginning of this process, each node selects itself as the steward and then progressively changes the local steward based on advertisements from the neighbors. StAR uses a sequence number to maintain the freshness of information, similar to AODV protocol.

Optimal Hop Count Routing (OHCR) and Minimum Power over Progress Routing (MPoPR): Hai, Amiya, and Ivan [74] are among the few researchers who leveraged controlled mobility of the robotic routers to assist in wireless data transmission among fixed nodes. This method is divided into two parts. The first, which they refer to as Optimal Hop Count Routing (OHCP), computes the optimal number of hops and optimal distances of adjacent nodes on the route. Each node identifies its closest node by comparing the respective neighbors' distances with the optimal distance. If a node cannot find any such neighboring node, it sends back a route failure message to the source. Otherwise, the second part of the routing, which the authors refer to as Minimum Power over Progress Routing (MPoPR), uses greedy routing on the results obtained from OHCP to minimize the total transmission power.

Synchronized QoS routing: In [75], Sugiyama, Tsujioka, and Murata presented a QoS routing technique for a robotic network that is based on the QoS routing in ad hoc network [76] and DSDV routing protocol [77]. In this method, they used the concept of Virtual Circuits to reserve a specified bandwidth. It is the job of the sender to reestablish the circuit in case of a broken connection due to topology changes. This method also includes a methodology for accelerating the transmissions of the control packets.

Topology Broadcast based on Reverse-Path Forwarding (TBRPF): In the CENTIBOT project [78], Konolige et al. used a proactive MANET technique called Topology Broadcast based on Reverse-Path Forwarding (TBRPF) to deal with multihop routing in dynamic robotic network. This link state routing protocol was originally proposed by Bellur and Ogiel [79, 80]. In this algorithm, each node maintains a partial source tree and reports part of this tree to its neighbor. To deal with mobility, it uses a combination of periodic and differential updates.

B.A.T.Mobile: Sliwa et al. [81] have also proposed a mobility-aware routing protocol called B.A.T.Mobile which builds upon the well-known B.A.T.M.A.N routing [64] protocol for MANET. This algorithm relies on a future position estimation module for the next hops that use the current and past position-related information as well as the knowledge of the mobility algorithms of the users. The estimation module is further used to rank the neighbors and estimate their lifetime. The neigh-

bor rankings are used to change the route in a proactive manner for end-to-end data transfer.

Other Methods: There also exist some methods that have the potential to be used in an RWSN after few modifications. Among such methods, the geographic routing algorithms and encounter-based routing algorithms are mentionable. Greedy Perimeter Stateless Routing (GPSR) [82] is an example of geographic routing protocols that use router positions and packet destinations for making forwarding decisions. If the locations of all nodes in the network are known, this algorithm can be used in RWSN. However, this approach faces many problems in mobile wireless networks. In [83] Son, Helmy and Krishnamachari identified two problems due to mobility in geographic routing, particularly in GPSR, called LLNK and LOOP. They also presented two solutions: neighbor location prediction (NLP) and destination location prediction (DLP); to solve those problems. Rao et al. [84] also identified some issues with GPSR and proposed a lifetime timer-based solution. In [85], Mauve et al. presented a generalized multicast version of GPSR like geographic routing. There are many other works on position-based routing [86, 87]. For a more detailed and complete overview on position-based routing algorithms, an interested reader is referred to [88].

Encounter-based routing is another relevant group of routing, mainly used in Delay Tolerant Networks (DTN). In general, DTN routing protocols are divided into two categories: forwarding-based or replication-based. Forwarding-based protocols use only one copy of the message in the entire network while the replication-based technique uses multiple copies of the message. Replica-based protocols are also subdivided into two categories: quota-based and flooding-based. Flooding is the most simple and inefficient technique. Balasubhamanian, Levine, and Venkataramani presented a flooding-based technique of replication routing in DTN [89, 90], modeling it as a resource allocation problem. Another flooding-based technique is presented in [91], called Maxprop. Spyropoulos, Psounis, and Raghavendra presented two quota-based replication routing techniques for DTN called Spray and Wait [92], and Spray and Focus [93]. There are many other papers on DTN routing [94–96]. Although this group of techniques is not directly related, they can be modified to develop very efficient routing for RWSN.

The research works related to data collection protocols in WSN community are also of interest. Among these protocols, a prominent and recent class of queue-aware routing algorithms, called Backpressure routing algorithms [97, 98], has caught our interest. The Backpressure routing algorithms and a range of similar algorithms [99–101] are proved to be "throughput optimal", in theory. One of the most recent Backpressure style routing algorithm is called the Heat Diffusion (HD) routing algorithm [102, 103] that has shown to offer a Pareto-optimal trade-off between routing cost and queue congestion (delay). The Backpressure routing algorithms, including HD algorithm, do not require any explicit path computations. Instead, the next hop for each packet depends on queue-differential weights that are functions of the local queue occupancy information and link state information at each node. There have been several reductions of Backpressure routing to practice in the form of distributed protocols, pragmatically implemented and empirically evaluated for different types

of wireless networks [98, 104–106]. Ghosh et al. have also developed a distributed practical version of the HD algorithm called Heat Diffusion collection protocol [107]. While these protocols perform effectively in a static WSN, their applicability in an RWSN are yet to be tested. Since these algorithms do not require any route calculation as well as routing tables, they will require less memory and computation in the resource-constrained robotic nodes. Moreover, one extra advantage of such protocols is the adaptability in a dynamic network due to not relying on a single predetermined path. Besides these protocols, there exist a number of other prior works on routing and collection protocols for wireless sensor networks, including the Collection Tree Protocol (CTP) [108], Glossy [109], Dozer [110], Low-power Wireless Bus [111], and RPL [112]. These protocols can also be modified for application in RWSN. The work presented in Glossy [109] is of particular interest due to its simplicity and wide adaptability for high throughputs and low delays. A concise overview of all the routing related work is presented in Table 3.

What Would Be the Potential Future Research Directions? To our knowledge, there exists a significant amount of research on routing related to MANET and WSN that can be applied to an RWSN either directly or after some modifications. However, a significant focus of future routing algorithms needs to be directed toward reducing delays, improving reliability, and incorporating the controlled mobility in the routing decisions. The emphasis should be on delay and reliability as on-time message delivery among different control system components is the key for a successful and efficient control system. One example of using the controlled mobility to our advantage is shown in the works of Wang, Gasparri, and Krishnamachari [113] where the robots ferry messages from a source to sink in a way similar to a postman. Another research direction would be to add node movements in routing decisions. For illustration, assume that there exist two possible routing paths, and the relatively bad path can be improved considerably by slightly moving the node. Then, the routing decision should include movement into consideration.

3.3 Connectivity Maintenance

Connectivity maintenance is a well studied and classic problem in the field of swarm robotics. In the connectivity maintenance problem, the main goal is to guarantee the existence of end-to-end paths between every pair of nodes. The interaction between pairs of robots is usually encoded by means of a graph, and the existence of an edge connecting a pair of vertexes represents the fact that two robots can exchange information either through sensing or communication capabilities. *Notably, the connectivity of the interaction graph represents a fundamental theoretical requirement for proving the convergence of distributed algorithms in a variety of tasks, ranging from distributed estimation* [114–116] *to distributed coordination and formation control* [117–119].

Traditional Approach: In the context of robotic networks, where the connectivity of the interaction graph is strictly related to the motion of the robots, a fundamental challenge is the design of distributed control algorithms which can guarantee that

Table 3 Summary of routing-related works

Routing algorithm name/class	References	Comments
MRSR, MRDV, and MRMM	[65]	Based on dynamic source routing (DSR) [66], AODV [67], and on-demand multicast routing protocol [68], respectively
Adaptive energy efficient routing protocol (AER)	[69]	Similar to DSR [66]
ACTor-based robots and equipment synthetic system (ACTRESS)	[70]	Main steps are negotiation, inquiry, offer, announcement, and synchronization
WNet	[71]	Based on optimized link state routing (OSLR) protocol [63]
Steward-assisted routing (StAR)	[73]	It is a hierarchical routing protocol where a group of robots acts as "stewards" that can store the data until a route to the destination is available
Optimal hop count routing (OHCR) and minimum power over progress routing (MPoPR)	[74]	Uses the controlled mobility of the robotic routers to assist in wireless data transmission among fixed nodes
Synchronized QoS routing	[75]	Based on the QoS routing in ad hoc networks [76] and DSDV routing protocol [77]
Topology broadcast based on Reverse-Path Forwarding (TBRPF)	[78]	It is a link state routing protocol. Each node maintains a partial tree
B.A.T.Mobile	[81]	Based on B.A.T.M.A.N routing [64]
Geographic routing algorithms	[82–88]	Employs locations of the nodes for efficient routing in WSN
Encounter-based routing		Routing algorithms for delay tolerant networks
Flooding-based	[89–91]	
Quota-Based	[92, 93]	
Data Collection routing in WSN		Data collection routing algorithms are used in WSNs for efficient routing of the sensed data to the data sinks

(continued)

Table 3 (continued)

Routing algorithm name/class	References	Comments
Backpressure routing	[97, 98]	
Heat diffusion routing	[102, 103, 107]	
Glossy	[109]	
Collection tree protocol	[108]	
Dozer	[110]	
Low-power wireless bus	[111]	
RPL	[112]	
Potential future directions	Modify existing routing protocol to include controlled mobility	
	Focus more on delay reduction, data transfer guarantee, and energy efficiency	

the relative motions of the robots do not result in a network partitioning, by relying only on local information exchange. Two possible versions of the connectivity maintenance problem can be considered: local connectivity and global connectivity. The local version of the connectivity maintenance problem focuses on the preservation of the original set of links of the graph encoding the pairwise robot-to-robot interactions to ensure its connectedness. The global version of the connectivity maintenance problem focuses on the preservation of the overall graph connectedness, i.e., links can be added or removed as long as this does not prevent the interaction graph to remain connected over time. Historically speaking, the local version of the connectivity problem has been the first version of the problem to be investigated by the research community. However, it turned out that preserving each of the links of the interaction graph significantly constrains the robots' mobility, while, in general, not each link is strictly required to ensure the connectedness of the interaction graph. For this reason, more recently the research community has focused mostly on the global version of the problem. Next, we present a brief survey of the available connectivity maintenance protocols that can be used in RWSN.

As already mentioned, connectivity maintenance has been studied extensively in the contexts of distributed robotics and swarm robotics. Most of the state-of-the-art protocols for connectivity maintenance are based on a graph modeling of the robot-to-robot interactions. More specifically, let $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ be the interaction graph where $\mathcal{V} = \{1, \ldots, n\}$ is the set of robots and $\mathcal{E}$ is the set of edges encoding the interactions between pairs of neighboring robots. In particular, the existence of an edge is often related to the spatial proximity between pairs of robots, i.e., an edge exists between two robots if the Euclidean distance between them is less than a given threshold.

By following this graph-based modeling of a robotics network, a natural metric to measure the network connectedness is the *algebraic connectivity*. More specifically, in the context of graph theory, the *algebraic connectivity* is defined as the second smallest Eigenvalue, $\lambda_2(\mathcal{L})$, of the graph Laplacian matrix, $\mathcal{L}$, of the network. In [120, 121], it is shown that $\lambda_2(\mathcal{L})$ is a concave function of the Laplacian matrix and represents the network connectivity when $\mathcal{L}$ is a positive definite matrix. Thus, the connectivity optimization goal becomes simple maximization of the algebraic connectivity value, $\lambda_2(\mathcal{L})$. Another way to represent the connectivity is via a powered sum of adjacency matrix, $\mathcal{A}_{sum} = \sum_{i=0}^{K} \mathcal{A}^i$ where $\mathcal{A}$ is the adjacency matrix of the network graph $\mathcal{G}$. $\mathcal{A}_{sum}$ basically represents the number of paths up to length K between every pair of nodes in the graph [122]. It follows that for a network to be connected, for all pairs of nodes and $K = n - 1$, $\mathcal{A}_{sum}$ has to be positive definite (where n is number of nodes).

Next, we briefly discuss some representative state-of-the-art local and global methods for connectivity maintenance. In [123], Dimarogonas and Kyriakopoulos presented one illustrative local connectivity maintenance approach using two potential fields in the controller where the nodes try to maintain all the initial links throughout the time. Notarstefano et al. also presented a double integrator disk graph-based local approach for connectivity maintenance [124, 125]. In [126], Spanos et al. introduced a concept of geometric connectivity robustness which is basically a function to model and optimize local connectivity. Another class of local connectivity-related works

lies in the context of leader–follower robot architectures where the follower robots try to maintain the connectivity to a designated leader or vice verse. The work of Yao and Gupta [127] is relevant in this context. They employed a leader–follower control architecture for connectivity by adaptively classifying the nodes into backbone and non-backbone nodes. Gaustavi et al. [128] have followed a similar path by identifying sufficient conditions for connectivity in a leader–follower architecture of mobile nodes. On the other hand, there exists a range of global connectivity-related works that are proposed over last decade. One of the earlier global decentralized connectivity maintenance techniques is the super-gradient and orthonormalization-based approach by Gennaro and Jadbabaie [129]. Later on, Dimarogonas and Johansson [130] proposed a control strategy using "bounded" inputs. Another very effective approach for connectivity maintenance based on decentralized estimation is presented in [33]. There are many extensions to this framework such as the integration of additional (bounded) control terms [131] and the saturation of the connectivity control term itself [132]. Zavlanos et al. also presented a couple of important techniques on the distributed global connectivity control [133–135] along with a compact survey on graph theoretic approaches for connectivity maintenance [136]. In [137], a technique based on dynamics of consensus methods is presented. Schuresko et al. [138] also presented techniques for connectivity control based on information dissemination algorithm, game theory, and the concept of spanning tree. A multi-hop information-based global connectivity maintenance and swarming technique is introduced by Manfredi [139]. Gil, Feldman, and Rus [140] proposed a well-known k-center problem-based connectivity maintenance algorithm for an application context where a group of robotic routers provides routing support to a set of robotic clients. The concept of bounded velocity of the routers and the clients is employed in this work. In [141], the connectivity maintenance problem in multi-robot systems with unicycle kinematics is addressed. In particular, by exploiting techniques from non-smooth analysis, a global connectivity maintenance technique under non-holonomic kinematics is proposed, which only requires intermittent estimation of algebraic connectivity, and accommodates discontinuous spatial interactions among robots. Most of these global connectivity maintenance methods are built upon the concepts of graph theory and algebraic connectivity.

Realistic Approach for RWSN: While all the previously mentioned methods for connectivity are relevant to the field of RWSN, most of them lack a communication channel model that includes the effect of fading and shadowing observed in a standard wireless channel. Rather, most of these methods employ the simple unit disk model for wireless communication links to model the network graph where every pair of nodes are assumed connected if and only if they are located within a communication radius, say R, of each other and disconnected otherwise. However, in reality, the communication links are very dynamic and unpredictable due to effects like fading and shadowing [5]. Therefore, the unit disk model-based connectivity maintenance are rather impractical and should be modified. Moreover, the connectivity maintenance algorithms should use an optimization function that takes into account communication link features such as signal strengths, data rates, realistic communication models, and line of sight maintenance to define connectedness. As an

example, Mostofi [142] presented a realistic communication model-based decentralized motion planning technique for connectivity maintenance. A behavioral approach for connectivity that takes into account the locations, measured signal strength and a map-based prediction of signal strengths is proposed by Powers and Balch [143]. Anderson et al. presented a line of sight connectivity maintenance technique via a network of relays and clusters of nodes in [144]. In [145], a spring–damper model-based connectivity maintenance is described. In summary, there exist only a handful of connectivity protocols that incorporate the well-known characteristics of an RF channel such as fading, and even fewer are practically implemented and evaluated (Table 4).

What Would Be the Potential Future Research Directions? One obvious future direction would be to extend the theory of traditional unit disk model-based connectivity maintenance protocol to include the effects of fading and shadowing. A modular or hierarchical approach would be ideal in this context where a graph identification module (by including fading and shadowing effects) and a graph theory-based connectivity maintenance module will work independently but with synergy. Another future direction would be to develop more protocols of second kind and evaluate them extensively to enrich the literature. Lastly, but most importantly, there is a lack of real-world experiments with a physical RWSN testbed to validate most of the existing theoretical and algorithmic contributions. Thus, a future goal of connectivity maintenance-related research should be on the development of a cheap, scalable, and easily programmable physical system and demonstration of the feasibility of the well-known solutions.

3.4 Communication-Aware Robot Positioning and Movement Control

In this section, we present a summary of the state-of-the-art techniques on communication-aware positioning and placement control of a group of robots in order to fulfill data routing and sensing requirements.

3.4.1 Multi-Robot Sensing

One of the main focus of RWSN should be on multiple robots-based sensing with realistic wireless communication constraints. Note that, every function/subproblem in an RWSN, such as localization and movement control, involves some sort of sensing such as RSSI, SINR, or locations of nodes. However, in this section, we focus on the state-of-the-arts on multi-robot systems where the main purpose of deployment is to sense an environment. There exist many works in the field of sensor networks and distributed robotics that deal with distributed sensing and sensed data collection. However, most of these works have some idealistic assumptions about either the communication model or the robot control problem and, thus, not directly applicable

Table 4 Summary of connectivity maintenance-related works

Importance	Exchange of information
	Proof of convergence
	Distributed coordination and formation control
	and so on
Versions	Local [123, 126, 127]
	Global [130, 137, 138, 140, 141]
Traditional approach	**Algebraic connectivity**:
	Second smallest eigenvalue, $\lambda_2(\mathcal{L})$, of graph Laplacian $\mathcal{L}$
	Represents connectivity if $\mathcal{L}$ is positive definite
	Maximize $\lambda_2(\mathcal{L})$ to improve connectivity
	Powered sum of adjacency matrix:
	$\mathcal{A}_{sum} = \sum_{i=0}^{K} \mathcal{A}^i$ where $\mathcal{A}$ is the adjacency matrix
	$\mathcal{A}_{sum}$ represents the number of paths up to length K between any pair of nodes in the graph
	Issues:
	Relies on simple unit disk model for interactions
	Lacks realistic communication channel model; effects of fading and shadowing [5] are ignored
	No focus on the communication link qualities
Realistic approach	**Features**:
	Accounts for location, signal strengths, interference, and data rates [142, 143]
	Realistic communication channel models [142]
	Line of sight maintenance between neighbors [144]
	Examples:
	[142–146]
Potential future directions	Modify existing unit disk model-based graph methods of connectivity
	Develop more efficient algorithm for connectivity with the focus on the realistic link qualities
	Develop hardware prototypes and test out the algorithm in real-world scenarios

in RWSN. In this section, we only focus on the existing works on multiple robots-based sensing that involves realistic models for both communication and control.

What Is Already Out There? In the field of multi-agent sensing, distributed coverage of the area of interest is a very well known topic of research in the contexts of both WSN [147] and multi-robot systems [148]. In the contexts of WSN, coverage control algorithms focus on the placements of static sensor nodes to optimally cover the area of interest. We do not present a survey of this well-studied problem. An interested reader is referred to a survey such as [147]. However, most of these works do not use the controlled mobility of the robots to the advantage. On the other hand, there exists a class of coverage control-related articles in the field of coordinated robotics that focus on the control and path planning of the robots. Most of these works employ graph theoretic tools such as Voronoi partitions to solve the coverage problem [149–151]. However, these works do not address the problem of collecting and communicating the sensed data effectively. Only recently, a small group of researchers started to look into multi-robot sensing problem from both control and communication point of views. The work of Kantaros and Zavlanos is relevant in this context [152]. They looked into the coverage problem of multiple robotic wireless sensors placement by formulating an optimization problem that combines placement optimization with realistic communication constraints and sensing efficiencies of the robotic nodes. In [153], Yan and Mostofi also looked into the problem of robotic path planning and optimal communication strategies in the context of a single robot-assisted sensing and data collections. Similar combined path planning and communication optimization in the contexts of multiple robots-based system are presented in the works of Ghaffarkhah and Mostofi [154, 155]. The works of Mostofi et al. [39, 156, 157] on cooperative sensing and structure mapping are also related to this context. In these works, the authors leveraged multiple pairs of coordinated robots and their RF communication abilities to sense/map unknown structures. To this end, they employed the concepts of compressible sampling/sensing [42] and the well-known propagation properties of RF signals such as path loss and fading [5]. The work by Le Ny, Ribeiro, and Pappas [158] also presents an optimization problem that couples motion planning and communication objective for sensing. On a similar note, Williams and Sukhatme proposed a multiple robot-based plume detection method in [159]. In [160], a new formulation of the multi-robot coverage problem is proposed. The novelty of this work is the introduction of a sensor network, which cooperates with the team of robots in order to provide coordination. The sensor network, taking advantage of its distributed nature, is responsible for both the construction of the path and for guiding the robots. The coverage of the environment is achieved by guaranteeing the reachability of the sensor nodes by the robots. A concise summary of this class of works is presented in Table 5.

What Would Be the Potential Future Research Directions? To our knowledge, there exist a very few works on robot-assisted sensing that involved timely, reliable, and efficient delivery of the sensed data. A major focus of the future should be on such joint optimization of data collection and sensing tasks. Moreover, there might exist many dissimilar robots (each consisting of different sets of sensors) in an RWSN. This requires a simple, unified abstraction in terms of control as well as the sensed data

Table 5 Summary of multi-robot sensing

Existing works	Multi-agent sensing and coverage algorithms from WSN [147, 148]
	Distributed coverage control in distributed robotics [149–151]
	Combined optimization of sensing coverage placement and efficient data routing and collection [39, 152–160, 162]
Potential future directions	Sensor data collection abstraction (e.g., publish–subscribe model) for multiple dissimilar robotic systems
	Sparse sensing for energy efficient nonredundant sensing
	More algorithms on combined optimization of robotic path planning, sensing, and communication quality

collection. The popular publish–subscribe-based frameworks (such as MQTT [161]) may be used in such contexts. Moreover, the introduction of controlled mobility has opened up the applications of sparse sensing [157] which can be exploited for energy efficient, nonredundant sensing.

3.4.2 Robotic Router Positioning

Robotic router/relay placement is a cutting-edge topic of research in the field of RWSN. It mostly concerns the second trend in RWSN research, i.e., the use of a robotic network to form a temporary communication backbone or support an existing backbone to improve performance. The problem of robotic router placements is complex and involves direct relations with many other research pieces of RWSN such as connectivity maintenance, communication link modeling, and localization. A robotic router/relay is a device with wireless communication capabilities and controlled mobility. Such devices can be employed to form temporary/adaptable communication paths and to ensure robust information flow between a set of nodes that wish to communicate but lack direct links between each other. Note that, we use "*robotic router/relay*" to refer to the robotic nodes helping in setting up communication and "*communication endpoints*" to refer to the nodes willing to communicate. Robotic relay/router nodes relay messages between such communication endpoints. The communication endpoints might have certain communication requirements such as minimum achievable data rate, high throughput, and lower delay. Moreover, the communication endpoints can be mobile or the environment can be dynamic with changing communication link properties. The objective of a robotic router placement/positioning algorithm is to place the relays in an optimal manner such that the communication requirements are fulfilled throughout the deployment time, and to adapt with the network dynamics. Before moving on, we present a commonly used

term in such contexts called "*flow*". *A "flow" is defined as the communication path between a pair of communication endpoints via a set of robotic routers/relays*. In other words, a set of robotic routers assigned to a flow are dedicated to form and support the communication path between the respective pair of communication endpoints. Based on the objectives as well as network dynamics, the allocation of a set of robotic routers among different flows may also change over time. Some of the major components of robotic router placements algorithms are as follows: proper positioning and movement planning of these robotic routers, allocation and reallocation of robots among flows as they arrive or disappear, and connectivity maintenance. In this section, we present the state-of-the-art techniques on robotic router placements.

What Is Already Out There? The earlier relevant state of the arts on robotic router placement/movement algorithms are linked to the connectivity maintenance problem. A major focus of such methods was to keep a moving target/node connected to a static base station via a set of robots with the assumption of an initially connected network. The work of Stump, Jadbabaie, and Kumar [163] is mentionable in this context where either the transmitter node is fixed and the receiver node is moving or vice versa. They developed a framework to control a team of robots to maintain connectivity between a sender and a receiver in such cases. Among other state of the arts, Tekdas, Yang, and Isler [164] focused on the connectivity of a single user to the base station and proposed two different models-based motion planning algorithms. One is based on known user motion (user trajectory algorithm) and the other is for unknown-random, adversarial motion of the user (adversarial user trajectory algorithm). However, this class of works does not deal with the qualities of the communication links as well as the end-to-end performance. De Hoog, Cameron, and Visser [165–167] have also proposed some techniques for maintaining connectivity to a command center in the context of exploration of unknown environments. In their tree-like role-based network formulation, the leaf nodes are the "explorers" that explore new frontiers, the root of the tree is the "base station", and the rest of the nodes are "relays" to keep connectivity between the "explorers" and the "base station".

Over last couple of years, a handful of researchers have started exploring the problem with more realistic communication models. Yan and Mostofi [168, 169] are among these handful of researchers to work on the robotic router problem. They extended the concept of connectivity maintenance to formulate an optimization problem which considers true reception quality expressed in terms of bit error rate. The goal was to minimize bit error rates of the receivers for two scenarios of multi-hop and diversity. *They also demonstrated that the Fiedler eigenvalue optimization-based approach results in a performance loss*. They used an extension of the channel modeling technique introduced in [45, 46]. In [170], Dixon and Frew presented a gradient-based mobility control algorithm for a team of relay robots with the goal of formation and maintenance of an optimal cascaded communication chain between endpoints. Rather than considering the relative positions of the neighbors, they used SNR of communication links between neighbors. They presented an adaptive extremum seeking (ES) algorithm which is employed for operating a distributed controller. Goldenberg

et al. [171] presented another distributed, self-adaptive technique for mobility control with the goal of improving communication performance of information flows. In their work, they tried to design and analyze a simple system to address three issues: application dependency, distributed nature, and self-organization. A solution to the problem of computing motion strategies for robotic routers in a simply connected polygon environment is presented in [172]. For a summary of all the relevant references, a reader is referred to Table 6.

As mentioned earlier, the main goals of robotic router placements are to fulfill certain communication requirements such as supporting a set of flows [173], guaranteeing certain performance criterion (say, data rate) to the customers [174], or fixing holes [175]. In [173], Williams, Gasparri, and Krishnamachari presented a hybrid architecture called INSPIRE, with two separate planes called Physical Control Plane (PCP) and Information Control Plane (ICP). Their goal was to support a flow-based network between multiple pairs of senders and receivers using a group of robots and optimize the overall Packet Reception Rate (PRR) (or the expected number of transmissions, ETX). In [113], Wang, Gasparri, and Krishnamachari presented a method called robotic message ferrying, where a set of robots literally travel from a source to a sink/destination to deliver data. The main objective of this work was the allocation of such robotic router nodes among a set of senders and the optimization of the communication performance such as throughput. Tuna et al. [175] also proposed a centralized method of fixing routing holes (due to the absence of nodes or failure of nodes) using a group of robotic routers. In the proposed method, all nodes communicate to a central server to send the sensed data, which in turn controls the positioning and deployment of a set of UAVs to fix routing holes. This algorithm employs geographical routing and Bellman–Ford routing algorithms to find the missing nodes/links. In [53], Fink, Ribeiro, and Kumar focused on guaranteeing a minimum end-to-end rate in a robotic wireless network. They model the communication channels via a Gaussian process model learned with a set of initially collected data. They proposed a stochastic routing variable to calculate an end-to-end rate estimate, which is exploited to find a slack in the achieved rate and the required rate. This estimated slack is further used for proper mobility control. Fida, Iqbal, and Ngo [176] proposed a new metric, called reception probability, and a throughput-based route optimization method that employs the new metric. They used Particle Swarm Optimization (PSO) [177] for finding the optimal configuration due to non-convexity of the problem. Gil et al. [174], also proposed a method of robotic router placements where the communication demands (in terms of data rates) of a set of clients are fulfilled by another set of robotic routers. The demands are modeled in terms of effective SNR (ESNR) to represent the required rate. Each client is serviced by the router closest to it while the router to router communications are assumed to have a very high capacity. They used a synthetic aperture radar (SAR) [178] concept-based directional signal strength (both amplitude and phase) estimation method and a Mahalanobis distance-based cost function for the positioning and path planning of the routers. Some preliminary works on optimizing SINR of the links, i.e., minimizing the effect of interference is presented in [179]. The work of Wang, Krishnamachari, and Ayanian [54] on robotic router placements in cluttered

Table 6 Summary of robotic router placements related works

Existing works	Tethering a moving object to a base station [163, 164]
	Robot-assisted static relay placements [181, 182]
	Single communication chain formation with performance guarantee in terms of bit error rate, number of hops, end-to-end rate, or SINR [53, 168–172, 176, 183–185]
	Multiple flow-based robotic network [173, 179] or mesh robotic network placement and performance optimization [174, 175]
	Use controlled mobility to carry the message from a source to a destination [113]
	Integrated framework for network goal-oriented mobility control [54, 173, 183]
Potential future directions	Hardware implementation and evaluation of the existing algorithms
	Include the effects of interference and channel access protocols in the placement optimization problem
	More decentralized method developments

environments, is also related to this context. On a related note, Ghosh and Krishnamachari [180] showed that there exists a bound on the number of robotic routers we need to deploy to guarantee certain communication requirements in terms of SINR. They also proposed a method of estimating worst-case interference and SINR in a flow-based robotic router deployment context.

What Would Be the Potential Future Research Directions? To our understanding, there exist a lot of research opportunities in this field of research. First, there is a lack of a physical robotic system-based experimentation of the existing works. One potential direction is to implement some of the promising algorithms and concepts on a real system and perform thorough analysis. To this end, there is a lack of academic open-source robotic network testbeds. Thus building a generic, scalable, adaptable robotic network testbed is another potential direction of research. Furthermore, most of the solutions are centralized and need to be converted to decentralized methods. Interference among the robotic routers as well as the effect of CSMA or any other channel access is also unaccounted for in most of the existing works. Nonetheless, to our understanding, the main focus of future should be on developing scalable, adaptable physical systems to test out the developed algorithms.

3.5 Localization

In this section, we present a survey on the state-of-the-art localization techniques in the context of RWSNs. The problem of localization is very well known in the contexts of sensor networks and distributed robotics. The state-of-the-arts on localization are very mature and require a complete separate survey [34–37]. In this section, we provide a very brief overview of the existing localization works and point out the works most relevant to the field of RWSN. Note that, the concept of "**localization**" is to locate a node in a deployment arena with respect to a reference frame or a reference location. A commonly used system called the Global Positioning System (GPS) localizes objects in terms of their latitudes and longitudes. However, GPS is known to not work properly in cluttered or indoor environments. Therefore, most of the target application contexts of RWSN require an alternate and efficient localization scheme for indoor environments such as RF-based localization. Moreover, while absolute locations are much important, a relative localization between the nodes in the network is sufficient in many contexts of RWSN. For example, consider a scenario where a group of robotic routers is employed to connect a moving target with a base station. In such contexts, the robots form a chain where each robot positions itself with respect to its neighboring nodes only. Relative positions with respect to the neighboring nodes are enough for a node's movement control decisions in this context. In this section, we present a summary of the existing literature on the relevant localization techniques.

Vision and Range-Based System: As mentioned earlier, localization has been a very active field of research in the domain of distributed robotics. The most popular localization systems in the field of robotics employ cameras and range finders. With the help of efficient sampling and filtering algorithms such as particle filtering or Kalman filtering, the camera-based systems locate the object in its field of view while range finders provide depth/distance information [186–189] . In order to deal with the movements as well as errors, some researchers use temporal snapshot of the targets [190, 191]. However, any camera/vision-based approach has many limitations such as the visibility requirement, limited field of vision of traditional cameras, larger form factor of the robots, and costly image processing software requirements. On the other hand, while the laser range-based methods do not suffer from the visibility problem, they are limited to direct line of sight between the target and the tracker and require complicated processing.

RF-Based System: As an alternative to GPS and vision-based localization systems, wireless sensor network researchers have proposed a variety of Radio Frequency (RF) based localization systems. As mentioned earlier, there exist a range of survey papers with the sole focus on such RF localization techniques [34, 35]. Briefly speaking, the existing localization techniques employ either of the following aspects: Direction of Arrival (DoA), Time of Arrival (ToA), Time Difference of Arrival (TDoA), Received Signal Strength (RSS), and proximity. The typical underlying technologies used to realize these techniques are RFID, WLAN, Bluetooth, and ZigBee. Liu et al. [34] provided a great outline and classifications of general

wireless localization techniques and systems. They outline most of the performance metrics used in traditional RF-based localization as follows: accuracy (mean error), precision (variance, or distribution of accuracy), complexity, robustness, scalability, and cost. With the recent trends of networked robots, it is of interest to the research community to look at these performance metrics from the perspectives of RWSN. The difference between an RWSN and a WSN with static and mobile actuating sensor nodes is that robots are more dynamic in position and require greater performance and flexibility in localization due to the larger range of tasks the robots are expected to perform.

RFID-Based System: Over last two decades, many researchers have peered into the use of RFID tags because of their *low cost and power (or zero power for passive tags)*. A confined deployment arena for a team of robots can be fitted with a mass deployment of RFID tags. Then, we can localize any RFID/RF device carrying robots in that arena [192, 193]. Zhou et al. present a survey of existing research and deployments of RFID tags for localization in [194] which shows its usefulness in robotics. However, one large quirk of RFID tags is the static nature of the tag placements and limited tag functions and information. The benefits of RFID tags manifest when meticulous planning or post-deployment positioning (using a robot) is done. In [195], a multi-robot exploration of an unknown graph describing a set of rooms connected by opaque passages is considered. In particular, the authors demonstrate how in this framework, which is appropriate for scenarios like indoor navigation or cave exploration, communication can be realized by bookkeeping devices, such as RFID tags, being dropped by the robots at explored vertices, the states of which are read and changed by further visiting robots. As an alternative, RFID tags can be replaced or complemented by WSNs, which have greater capabilities and flexibility.

Wireless AP-Based System: With the ubiquity of WLAN access points and wide availability of wireless sensor nodes, the research community has investigated the use of a network infrastructure to position and navigate robots. The work of Ladd et al. [196] illustrated the feasibility of using commercial off-the-shelf radios and radio signal strength measurements as a robust locater of robots. To enhance the overall performance of RSS-based localization, researchers have investigated RF scene analysis or fingerprinting as a viable option for indoors. Ocana et al. [197] presented such a robot localization system that starts with a semiautonomous method to fingerprint indoor environments using a robot. Many other research teams followed up with works concentrated on RF surveying with directional antennas on a robot, such as in [198–202]. We encourage readers to also refer to our section on RF mapping (Sect. 3.1) to complement the RF scene analysis.

Distributed and Cooperative System: Distributed and cooperative network and RF-based localization in a dynamically moving network of robots is still not a fully understood area. Wymeersch [203] showed the positive impacts of cooperative localization on achieving more complex tasks with a team of robots. This further motivates the need for a better understanding of cooperative network and RF-based localization to see if we can enhance the functionality of RWSNs. In [204], Koutsonikolas et al. delved into the problem of cooperative localization, but the authors assume there is a subset of robots that carry extra sensors, including GPS, to aid in the

overall system. The rest of the robots carry 802.11 radios, only using beacons to determine proximity. On that node, the work of Zickler and Veloso [205] focus on relative localization and tethering between two robots based on the received signal strengths. They opt for a discrete grid-based Bayesian probabilistic approach. In their system, a locator node moves to different relative positions with respect to the node being localized to collect multiple RSS values while they communicate their odometer readings. In [206], Filoramo et al. describe an RSSI-based technique for inter-distance computation in multi-robot systems. In particular, for a team of robots equipped with Zigbee radio transceivers, they propose a data acquisition technique which relies on spatial and frequency averaging to reduce the effect of multipath for both indoor and outdoor environments. Furthermore, they show how the proposed data acquisition technique can be used to improve a Kalman Filter-based localization approach. Another RSSI-based relative localization system is proposed by Oliveira et al. [207] which also relies on pairwise RSSI measurements between the robots. To improve the performance and accuracy of the localization, they apply Kalman filter and the FloydWarshall algorithm. One of the most recent significant works on relative localization is presented in [208] which applies MIMO-based system for a single node-based localization. These methods are particularly relevant to the field of RWSN. We present a summary of the state-of-the-art localization methods in Table 7.

What Would Be the Potential Future Research Directions? Most of the current robotics network-related researches are performed in artificial fixed indoor environments as there exist costly camera-based solutions (such as VICON system [209]) to provide millimeter-level accuracy required for the experimentation. We believe that RF localization will be able to help extend such experiments to truly indoor cluttered environments with much lower cost than deploying camera systems. However, most of the existing RF solutions are also not portable, e.g., they require either a fixed infrastructure with RF beacons, a map of the environment, etc. Therefore, the future direction of localization relation research should be focused on developing **portable scalable** at least centimeter-level accurate RF localization systems that can be quickly deployed on demand and can be removed easily. Moreover, the frameworks should be portable to deploy in real-world applications such as in firefighting. In that context, another potential direction is toward "self-sufficient" RWSN, i.e., the network will not require help from any existing infrastructure to perform the assigned tasks. Further, researchers should study different localization properties such as accuracy, complexity, and communication requirements in the context of RWSN where a robot's performance directly relies on localization, e.g., in the context of tracking and following a firefighter while providing connectivity to a command center.

4 RWSN Network Stack Layer Analysis

In the last section, we categorized the state-of-the-arts in the field of RWSN according to the key research problems in focus. The majority of the works in RWSN

Table 7 Summary of localization-related works

Localization algorithm type/class	References	Comments
Vision and range-based system	[186–191]	Popular in robotics
		High accuracy
		Many off-the-shelf solutions are available
		Requires heavy computation
		Requires line of sight
		Large form factor
RF-based system	[34, 35]	Popular in WSN community
		Uses RF signal properties (like signal strength or time of arrival)
		Lower accuracy compared to camera-based system
		Low cost and low computation
		No line of sight requirement
		Small form factor
RFID-based system	[192–195]	Low cost and power (or zero power for passive tags)
		Requires static placements of the tags in the deployment arena
		Requires pre-deployment of the localization infrastructure
Wireless AP-based system	[196–202]	Use pre-deployed or existing WLAN access points
		Use commercial off-the-shelf radios
		Requires pre-mapping or fingerprinting of the deployment region
Distributed and cooperative system	[203–208]	A subset of anchor robots has GPS or other localization capabilities
		Rest of the robots localize themselves relative to the anchor nodes
		Sometimes odometer readings are combined for better accuracy [205]
		Some types of filtering can be involved to improve performance

(continued)

Table 7 (continued)

Localization algorithm type/class	References	Comments
Future directions	Co-optimization of localization accuracy and communication goal	
	Develop cheap, scalable, and high accuracy system	
	Study the trade-offs among accuracy, complexity, and cost	

are focused on *RSS modeling and mapping*, *connectivity maintenance*, *routing algorithms*, *communication-aware robot placements*, *connectivity maintenance*, and *localization*. However, one key thing missing in the last section is how all these works fit in the contexts of traditional networking concepts. Traditionally, the protocols relevant to a network (say sensor network) are developed by following the well-known layered network stack architectures such as the OSI model or Internet model. Ideally, we would want the same for RWSN. However, we discover that the layered structure in an RWSN device does not follow the traditional norm. Rather, it mostly relies on interdependencies between layers. To understand the layering requirements, we first analyze the existing works explained in the last section according to the five-layered Internet protocol stack: physical layer, MAC layer, network layer, transport layer, and application layer. Next, we discuss some potential unified architectures for RWSN.

4.1 Internet Model for Network

4.1.1 Physical Layer

The physical layer in a traditional network deals with the physical communication between nodes. The function of physical layer includes but not limited to representation of bits, controlling data rate, synchronization between transmitter and receiver, defining the communication interface, and controlling the mode of communication such as simplex or duplex. Therefore, research on topics such as communication hardware, physical medium and communication technologies (e.g., RF and Bluetooth), and modulation–demodulation of signal falls under this category. There exist many sorts of wireless communication technologies such as RF, Bluetooth, and Sonar for communication among robots. For obvious reasons, the dominant technologies for aboveground communication are RF communication techniques [210–213]. For short range communication, some networks of robots use Bluetooth [214] and infrared. The aboveground radio frequency-based methods are not applicable for acoustic communication (e.g., underwater communication) due to reasons like high loss exponents and fading due to turbulence. The most common acoustic communication techniques are special RF communication [215] and sonar. While RF-based communication still remains the mainstream for RWSN system, the introduction of robots and controlled mobility has opened up some new but relatively unexplored communication methods. On that node, Ghosh et al. [216] have presented a proof-of-concept of a new

method of communication that employs the location and the motion pattern of a communicating robot as the communication signal.

Most of the works on RSS measurements, RF mapping, and position control to improve link qualities fall under the purview of the physical layer since these works are directly linked to the signal strengths variations over the area of interest. On a similar note, most of the RF-based localization techniques use properties (such as signal strength) of the communication signals (RF or ultrasound) and, thus, also fall under the purview of the physical layer. Among the emerging research domains related to physical layer of an RWSN, the concept of distributed MIMO implementation using robots is promising. In [217], Zhang et al. presented a cooperative MIMO like communication structure in mobile sensor/robotic networks. They subdivide a network into twin subnetworks where each transmitter node pairs itself with another node for transmitting cooperatively in an MIMO like fashion.

Another class of work that also partly falls under the purview of physical layer lies within the communication-aware robot positioning and movement control-related works where the robots adapt their positions to optimize the quality of the communication links. In such cases, the robots employ the physical layer information such as RSSI [179] or data rate [174] to control the movements of the robots acting as relays/routers/sensors.

In summary, we can state that a major focus of most of the existing state of the arts on RWSN has been toward the physical layer.

4.1.2 Media Access Control Layer

Briefly speaking, Media Access Control (MAC) protocols deal with proper distributed access of the physical medium among nodes, node to node communication, framing, and error corrections. While most of the classical MAC layer modules and protocols [5] are applicable to RWSN, to our knowledge, there exist only a few MAC layer protocols designed specifically for networking between robots. Related to media access protocols, CSMA/CA is an obvious choice for RWSN due to its ubiquitous properties such as randomness and scalability. On the other hands, TDMA and FDMA systems can also be modified for RWSN with the extra feature of controlled mobility. On that note, Hollinger et al. [218] presented an MAC protocol for robotic sensor networks in acoustic environments. They proposed a three-phased method based on TDMA with acknowledgments. The phases are Initiation, Scheduling, and Data transfer, respectively. There also exist works related to mobile WSN which use predicted mobility patterns, such as pedestrian mobility or vehicular mobility, of both source and sink to design efficient application context specific MAC protocols. Some examples of such MAC protocols are as follows: MS-MAC [219], M-MAC [220], M-TDMA [221], MA-MAC [222], MobiSense [223], and MCMAC [224]. These works show that many difficulties arise as a result of mobility (mainly uncontrolled mobility) such as random variations in link quality and frequent route changes. To deal with these problems of mobility, researchers proposed a class of methods like negotiation-based rate adaptation and handover by transmitter. For a more detailed

survey on such techniques, the reader is referred to [225]. In contrast, the mobility of the nodes in RWSN are controlled and, thus, can be exactly known or predicted with higher accuracy. This opens up a new domain of research where the mobility controller and the MAC protocol could work in an integral manner to optimize the utilization of the radio resources.

In summary, while the existing MAC protocols are applicable to RWSN, there is a lack of MAC layer protocols specifically designed and optimized for RWSN. As mentioned earlier, one of the future directions would be to incorporate mobility control with MAC access protocols for better performance. Another future direction would be to use the knowledge of the MAC protocol to estimate signal properties such as interference and SINR [180], and to control the link properties such as interference [226].

4.1.3 Network Layer

Network layer in a traditional wireless network deals with the packetization of data as well as routing of the packets from source nodes to respective destination nodes. One of the main application contexts of RWSN is to maintain a temporary communication backbone to support data flow between communication endpoints. Thus, a major focus of RWSN-related works till date has been on developing network layer protocols and combining controlled mobility with routing of packets. All the routing-related works presented in Sect. 3.2 directly fall under this category for obvious reasons. Similarly, all the works on communication-aware robot placements (discussed in Sect. 3.4) also fall under the purview of network layer. However, as mentioned earlier, the concept of layering in RWSN is slightly vague as it relies on cross-layer dependencies such as dependencies between the physical layer and the network layer. This is most apparent in the contexts of robotic router placement algorithms that deal with the placements of the robotic nodes (based on the physical layer information) to optimize the end-to-end path from a source to destination as well as to optimize each link [81, 173]. All connectivity-related works presented in Sect. 3.3 also fall under this category as the key goal in such connectivity maintenance algorithms is to guarantee the existence of a communication path between every pair of nodes in the network. Without connectivity, the network might be segregated into smaller subnetworks and would not be able to fulfill the routing goals.

4.1.4 Transport Layer

In the field of RWSN, the researchers are yet to significantly focus on the transport layer protocols. Till date, researchers employed traditional transport layer protocols such as TCP, UDP, or some MANET transport layer protocols for robotic networks. However, unlike MANET, robotic networks have an extra feature of controllable mobility which provides an extra dimension. But there is no significant work on controlling mobility for improved performance of transport layer protocols. Conversely,

controllability requires highly reliable, low delay, and error-free communication between robots, which is not possible using original TCP or UDP and requires some special transport layer protocol. In this section, we present a brief overview of the existing transport layer protocols for RWSN.

Among the state-of-the-arts, the work of Douglas W. Gage on the MSSMP Transport Layer Protocol (MTP) [210] is mentionable. This protocol is based on the Reliable Data Protocol (RDP). The RDP is much more effective and appropriate service model for mobile robot applications than TCP. There are many features of RDP which are more useful in RWSN such as more communications bandwidth than TCP and simpler in terms of implementation. But the interface to the application layer in MTP is similar to TCP, i.e., based on socket-like API. There is also a couple of congestion control protocol designed for tele-operation of robots such as trinomial method [227], Real-Time Network Protocol (RTNP) [228], and Interactive Real-Time Protocol (IRTP) [229]. All these methods are not directly related to the RWSN but can be ported. Similarly, there exists a group of works on transport layer for MANET [230, 231]. For a detailed overview on existing congestion control protocols of mobile ad hoc network, an interested reader is referred to [232]. In summary, the transport layer-related research on RWSN requires significant attention in future with a major focus on reliability and delay performances.

4.1.5 Application Layer

Many of the existing works related to RWSN are actually related to the application layer. In some of the target application contexts, the robots in an RWSN need to make informed movement and communication decisions in order to work in a cooperative manner. For example, in order to form a communication relay path between a pair of communication endpoints, the relay robots need to process a combination of information such as neighboring node locations, flow endpoint requirements (say, a minimum data rate), current link status, and expected interference power to adaptably and optimally position the relays. According to the layered hierarchy, all these processing and decision-making should be done in the application layer to keep the system modular. For example, in the route swarm work [173], the Information Control Plane (ICP) takes care of making decision regarding state changes of the robots as well as the allocation of robots among different flows. Thus, the ICP in that architecture is mostly implemented in the application layer. In the same manner, all the robotic router-related works are partly/fully dependent upon the application layer.

Another mentionable field related to the application layer is cloud robotics. Cloud robotics basically uses an application layer abstraction of a heterogeneous network of robots to perform a group of tasks. Cloud robotics provides a unified scalable control platform for a group of heterogeneous robots. The work of Du et al. [29] is among the most promising works in this field of research. Quintas et al. [30] also proposed a related architecture. In [31], Kamei et al. presented a detailed study of the advantages, concerns, and feasibility in Cloud Networked Robotics. In this paper, we

Table 8 Summary of relevant keywords for the RWSN layering architecture

Layer	References	Related keywords
Physical layer	[210–217]	RSS measurement, RF mapping, localization, distributed MIMO, connectivity, robotic router placement, and communication-aware robot positioning
MAC layer	[5, 180, 218–226]	Scheduling, CSMA, TDMA, and FDMA
Network layer		Routing, robotic router placement, communication-aware robot positioning, and connectivity
Transport layer	[210, 227–232]	Delay, real-time communication, UDP, and TCP
Application later	[29–31, 173]	Connectivity, positioning, robotic router placement, communication-aware robot positioning, and cloud robotics

do not delve into cloud robotics as it does not directly fall under the RWSN research domain.

In summary, there is no clean way of classifying the existing works into the layered architecture. Rather, each of the problems and solutions belongs to multiple layers. Moreover, we need a new layer/module to deal with the mobility control. All these lead us to believe that maybe we need a new architecture for RWSN that builds upon the existing layered network stacks, discussed in the next section. Note that we present a summary of our network protocol stack-related discussion in Table 8.

4.2 An Unified System Architecture for RWSN

Based on our analysis in Sect. 4.1, we find that the existing works in RWSN do not fit well in the Internet model of networking stack. Rather, most of the solutions in RWSN require interlayer dependencies. For example, a robotic router placement algorithm relies on the physical layer estimation models which in turn rely on some knowledge about the relay node positions and the network graphs. Moreover, we need to have a control layer to combine the network goals with the movement of the robots. Thus, we require a new system hierarchy for RWSN where the existing network architecture can be kept intact to the most extent. On that note, Williams, Gasparri, and Krishnamachari [173] have proposed an architecture with two planes:

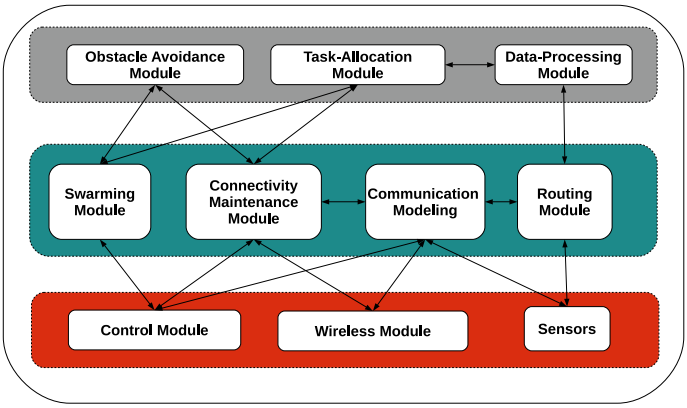


Fig. 4 Illustration of a unified architecture

the Information Control Plane (ICP) and the Physical Control Place (PCP) where the ICP takes care of the networking as well as high-level movement decisions and the PCP takes care of the movements. There also exist software architecture solutions for a system of multiple robots such as ALLIANCE [233] and CLARAty [234]. The works of Arkin and Balch [235] and Stoeter et al. [236] are also relevant in this context.

However, the concept of a unified system architecture is one of the relatively unexplored domains of research in RWSN. Most of the existing literature emphasize only certain aspects of system challenges instead of focusing on the networked robot system as a whole. Thus, there is a need of a **base, decentralized, realistic system framework using realistic communication models** that can autonomously control an individual robot as well as a group of robots in any kind of RWSN application contexts. **The term *framework* here refers to a collective set of system modules such as movement control, sensors and connectivity maintenance with necessary interconnections, as illustrated in Fig**. 4. This base framework should have **plug** & **play** flexibility as well, i.e., any extra module pertinent to a specific requirement can be added or removed. In Fig. 4, we present a sample, illustrative architecture for RWSN, based on our understanding.

5 Collaborative Works on Networked Robots

There are many projects on collaborative robotics with different goals in focus. Among them, Ubiquitous Robotics network system for Urban Settings (URUS) project (http://urus.upc.es) [237], Japan's NRS project [238], Physically Embedded Intelligent Systems Ecology [239] project, DARPA LANdroids program [240], Mobile Autonomous Robot Systems (MARS) [241], Mobile Detection Assessment

and Response System (MDARS) [210] are the important ones. Among other projects, the swarm-bot project by cole Polytechnique Fdrale de Lausanne (EPFL) ([242], [243]), the NECTAR Project [244] by Filippo Arrichiello and Andrea Gasparri, and I-Swarm project ([245], [246]) by Karlsruhe are the significant ones.

Since we are mainly interested in network-related research in RWSN, the main project that falls in our category is the DARPA LANdroids program. This is one of the recent projects undertaken on RWSN. Tactical communication enhancement in urban environments is the main goal of this program [240]. Toward this goal, the researchers tried to develop pocket-sized intelligent autonomous robotic radio relay nodes that are inexpensive. One of the serious communications problems in urban settings is multipath effect. LANdroids are envisioned to mitigate the problem by acting as relay nodes, using autonomous movements and intelligent control algorithms. LANdroids will also be used to maintain network connectivity between dismounted war-fighters and higher command by taking advantage of their cooperative movements.

On the other hand, there are some industrial projects that also fall under the purview of RWSN such as Facebook's Aquila project [247] and Google's project Loon [248]. There are also some open-source projects on swarm robotics such as swarm robot [249] and swarm-bots [250]. Swarming Micro Air Vehicle Network (SMAVNET) is a related project by EPFL where a swarm of UAVs is envisioned to be used to create temporary communication networks.

6 Summary and Conclusion

The main aim of this chapter was to identify and define a new field of research, RWSN, and provide a starting point to the new researchers. Briefly speaking, an RWSN consists of a group of controllable robots with wireless capabilities that are deployed with the goal of improving/providing a portable wireless network infrastructure in application with need of sudden and temporary wireless connectivity such as in a search and rescue mission or in a carnival. While there exist a range of relevant state-of-the-arts, the application of controlled mobility to the advantage of wireless communication is still an open area of research. However, like every new field of research, there are some challenges in RWSN research. Some of the known challenges are as follows: (1) lack of programmable, scalable RWSN testbeds for implementing and validating concepts; (2) lack of good venues to publish research (there is only a handful of new workshops and conferences that focus on RWSN); and (3) because of the interdisciplinary nature of this field, it requires the researchers to have knowledge on a vast range of topics such as robotics, control, communication, embedded systems, and networks. Nonetheless, based on all the discussions in this chapter, it is evident that RWSN is an emerging and promising piece of technology with limitless possibilities. Some of the known promising ongoing and future directions of RWSN-related research are listed in Table 9.

Table 9 Ongoing and future research directions

Systems	☐ Build a full-fledged low-power reusable RWSN testbed
	☐ Implement and analyze promising theoretical concepts on a real system
	☐ Extensive measurements in real environments (such as mines), identify the RF properties, and formulate communication models and emulators
Modeling and mapping	☐ Build a mathematical or systemic model for interference and SINR estimation in an RWSN
	☐ Incorporate the effects of MAC protocols such as CSMA into interference estimations
	☐ RF-based online mapping of an unknown environment
Routing	☐ Develop routing algorithms with guaranteed lower delay but higher reliability
	☐ Apply existing MANET protocols in the context of RWSN
	☐ Include controllability of the nodes in the routing decision where a bad link can be potentially improved via small movements
Connectivity maintenance	☐ Realistic communication model-based connectivity control
Robotic router	☐ Optimization of router placements with realistic SINR models
	☐ Guaranteed communication performance such as min achievable data rate by placing robotic routers between TX-RX pairs
	☐ Robot-based communication link repair
	☐ Robotic message ferrying-related research with more focus on the trade-off between movement energy consumption cost and the payoff from good performance or timely message delivery
Localization	☐ Build a portable RF-based localization system with at least centimeter-level accuracy
	☐ Focus more on relative localization than absolute localization
Network stack	☐ Multiple robots-based cooperative MIMO
	☐ MAC protocols with mobility control, engineered specifically for RWSN
	☐ Transport layer protocols (alternate to UDP or TCP) engineered for RWSN
	☐ Unified system architecture for RWSN

References

1. Penders, J., Alboul, L., Witkowski, U., Naghsh, A., Saez-Pons, J., Herbrechtsmeier, S., El-Habbal, M.: A robot swarm assisting a human fire-fighter. Adv. Robot. **25**(1–2), 93–117 (2011)
2. Murphy, R.R.: Trial by fire [rescue robots]. IEEE Robot. Autom. Mag. **11**(3), 50–61 (2004)
3. Gazi, V., Kevin, M.P.: Swarm Stability and Optimization. Springer (2011)
4. Aahin, E., Winfield, A.: Special issue on swarm robotics. Swarm Intell. **2**(2–4), 69–72 (2008)
5. Theodore, S.R.: Wireless Communications: Principles and Practice. Prentice Hall PTR, New Jersey (1996)
6. Ollero, A., Alcázar, J., Cuesta, F., López-Pichaco, F., Nogales, C.: Helicopter teleoperation for aerial monitoring in the comets multi-uav system. In: Proceedings of the 3rd IARP Workshop on Service, Assistive and Personal Robots (2003)
7. Thrun, S., Thayer, S., Whittaker, W., Baker, C., Burgard, W., Ferguson, D., Hahnel, D., Montemerlo, D., Morris, A., Omohundro, Z.: Autonomous exploration and mapping of abandoned mines. IEEE Robot. Autom. Mag. **11**(4), 79–91 (2004)
8. Murphy, R.R., Kravitz, J., Samuel, L.S., Shoureshi, R.: Mobile robots in mine rescue and recovery. IEEE Robot. Autom. Mag. **16**(2), 91–103 (2009)
9. Weiss, M.D., Peak, J., Schwengler, T.: A statistical radio range model for a robot manet in a subterranean mine. IEEE Trans. Veh. Technol. **57**(5), 2658–2666 (2008)
10. Baber, J., Kolodko, J., Noel, T., Parent, M., Vlacic, L.: Cooperative autonomous driving: intelligent vehicles sharing city roads. IEEE Robot. Autom. Mag. **12**(1), 44–49 (2005)
11. Nagel, R., Eichler, S., Eberspacher, J.: Intelligent wireless communication for future autonomous and cognitive automobiles. In: Proceedings of the IEEE Intelligent Vehicles Symposium (IV) (2007)
12. Milanés, V., Alonso, J., Bouraoui, L., Ploeg, J.: Cooperative maneuvering in close environments among cybercars and dual-mode cars. IEEE Trans. Intell. Transp. Syst. **12**(1), 15–24 (2011)
13. Xiong, N., Vasilakos, A.V., Yang, L.T., Pedrycz, W., Zhang, Y., Li, Y.: A resilient and scalable flocking scheme in autonomous vehicular networks. Mob. Netw. Appl. **15**(1), 126–136 (2010)
14. Parker, L.E.: Alliance: an architecture for fault tolerant multirobot cooperation. IEEE Trans. Robot. Autom. **14**(2), 220–240 (1998)
15. Ibach, P., Milanovic, N., Richling, J., Stantchev, V., Wiesner, A., Malek, M.: Cero: Ce robots community. IEE Proc.-Softw. **152**(5), 210–214 (2005)
16. Kovács, T., Pásztor, A., Istenes, Z.: Connectivity in a wireless network of mobile robots doing a searching and collecting task. In: Proceedings of the IEEE International Symposium on Applied Computational Intelligence and Informatics (SACI) (2009)
17. Steels, L.: Cooperation between distributed agents through self-organisation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (1990)
18. Stilwell, D.J., Bay, J.S.: Toward the development of a material transport system using swarms of ant-like robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (1993)
19. Nguyen, H.G., Pezeshkian, N., Raymond, M., Gupta, A., Spector, J.M.: Autonomous communication relays for tactical robots. Technical report, DTIC Document (2003)
20. Hsieh, M.A., Cowley, A., Keller, J.F., Chaimowicz, L., Grocholsky, B., Kumar, V., Taylor, C.J., Endo, Y., Arkin, R.C., Jung, B.: Adaptive teams of autonomous aerial and ground robots for situational awareness. J. Field Robot. **24**(11–12), 991–1014 (2007)
21. Erickson, J.K.: Living the dream-an overview of the mars exploration project. IEEE Robot. Autom. Mag. **13**(2), 12–18 (2006)
22. Calkins, D.: An overview of robogames [competitions]. IEEE Robot. Autom. Mag. **18**(1), 14–15 (2011)
23. Petelin, J.B., Nelson, M.E., Goodman, J.: Deployment and early experience with remote-presence patient care in a community hospital. Surg. Endosc. **21**(1), 53–56 (2007)

24. Baeg, S.-H., Park, J.-H., Koh, J., Park, K.-W., Baeg, M.-H.: Robomaidhome: a sensor network-based smart home environment for service robots. In: Proceedings of the IEEE International Symposium on Robot and Human interactive Communication (RO-MAN) (2007)
25. Baeg, S.-H., Park, J.-H., Koh, J., Park, K.-W., Baeg, M.-H.: Building a smart home environment for service robots based on rfid and sensor networks. In: Proceedings of the IEEE International Conference on Control, Automation and Systems (ICCAS) (2007)
26. De La Pinta, J.R., Maestre, J.M., Camacho, E.F., Alonso, I.G.: Robots in the smart home: a project towards interoperability. Int. J. Ad Hoc Ubiquitous Comput. **7**(3), 192–201 (2011)
27. Correll, N., Bachrach, J., Vickery, D., Rus, D.: Ad-hoc wireless network coverage with networked robots that cannot localize. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2009)
28. Batalin, M.A., Sukhatme, G.S.: Coverage, exploration and deployment by a mobile robot and communication network. Telecommun. Syst. **26**(2–4), 181–196 (2004)
29. Du, Z., Yang, W., Chen, Y., Sun, X., Wang, X., Xu, C.: Design of a robot cloud center. In: Proceedings of the International Symposium on Autonomous Decentralized Systems (ISADS) (2011)
30. Quintas, J., Menezes, P., Dias, J.: Cloud robotics: towards context aware robotic networks. In: Proceedings of the 16th IASTED International Conference on Robotic (2011)
31. Kamei, K., Nishio, S., Hagita, N., Sato, M.: Cloud networked robotics. IEEE Netw. **26**(3), 28–34 (2012)
32. Bullo, F., Cortes, J., Martinez, S.: Distributed control of robotic networks. Applied Mathematics Series. Princeton University Press (2009). http://coordinationbook.info
33. Yang, P., Freeman, R.A., Gordon, G.J., Lynch, K.M., Srinivasa, S.S., Sukthankar, R.: Decentralized estimation and control of graph connectivity for mobile sensor networks. Automatica **46**(2), 390–396 (2010)
34. Liu, H., Darabi, H., Banerjee, P., Liu, J.: Survey of wireless indoor positioning techniques and systems. IEEE Trans. Syst. Man Cybern. Part C: Appl. Rev. **37**(6), 1067–1080 (2007)
35. Sun, G., Jie, C., Wei, G., Liu, R.K.J.: Signal processing techniques in network-aided positioning: a survey of state-of-the-art positioning designs. IEEE Signal Process. Mag. **22**(4), 12–23 (2005)
36. Guvenc, I., Chong, C.-C.: A survey on toa based wireless localization and nlos mitigation techniques. IEEE Commun. Surv. Tutor. **11**(3), 107–124 (2009)
37. Amundson, I., Koutsoukos, X.D.: A survey on localization for mobile wireless sensor networks. In: Mobile Entity Localization and Tracking in GPS-less Environnments, pp. 235–254. Springer (2009)
38. Mostofi, Y., Sen, P.: Compressive cooperative mapping in mobile networks. In: Proceedings of the American Control Conference (ACC) (2009)
39. Gonzalez-Ruiz, A., Ghaffarkhah, A., Mostofi, Y.: An integrated framework for obstacle mapping with see-through capabilities using laser and wireless channel measurements. IEEE Sens. J. **14**(1), 25–38 (2014)
40. Gonzalez-Ruiz, A., Ghaffarkhah, A., Mostofi, Y.: A comprehensive overview and characterization of wireless channels for networked robotic and control systems. J. Robot. **2011** (2012)
41. Mostofi, Y., Sen, P: Compressed mapping of communication signal strength. In: Proceedings of the Military Communications Conference (MILCOM) (2008)
42. Candès, E.J.:: Compressive sampling. In: Proceedings of the International Congress of Mathematicians, vol. 3, pp. 1433–1452. Madrid, Spain (2006)
43. Mostofi, Y., Gonzalez-Ruiz, A., Gaffarkhah, A., Li, D.: Characterization and modeling of wireless channels for networked robotic and control systems-a comprehensive overview. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2009)
44. Mostofi, Y.: Compressive cooperative sensing and mapping in mobile networks. IEEE Trans. Mob. Comput. **10**(12), 1769–1784 (2011)

45. Malmirchegini, M., Mostofi, Y.: On the spatial predictability of communication channels. IEEE Trans. Wirel. Commun. **11**(3), 964–978 (2012)
46. Mostofi, Y., Malmirchegini, M., Ghaffarkhah, A.: Estimation of communication signal strength in robotic networks. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2010)
47. Yan, Y., Mostofi, Y.: Co-optimization of communication and motion planning of a robotic operation under resource constraints and in fading environments. IEEE Trans. Wirel. Commun. **12**(4), 1562–1572 (2013)
48. Mostofi, Y.: Cooperative wireless-based obstacle/object mapping and see-through capabilities in robotic networks. IEEE Trans. Mob, Comput (2012)
49. Depatla, S., Buckland, L., Mostofi, Y.: X-Ray vision with only WiFi power measurements using Rytov wave models. IEEE Trans. Veh. Technol. **64**, 1376–1387 (2015)
50. Hsieh, M.-Y.A., Kumar, V., Taylor, C.J.: Constructing radio signal strength maps with multiple robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2004)
51. Hsieh, A.M., Cowley, A., Kumar, V., Taylor, C.J.: Maintaining network connectivity and performance in robot teams. J. Field Robot. **25**(1–2), 111–131 (2008)
52. Fink, J., Kumar, V.: Online methods for radio signal mapping with mobile robots. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2010)
53. Fink, J., Ribeiro, A., Kumar, V.: Robust control of mobility and communications in autonomous robot teams. IEEE Access **1**, 290–309 (2013)
54. Wang, S., Krishnamachari, B., Ayanian, N: The optimism principle: a unified framework for optimal robotic network deployment in an unknown obstructed environment. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2015)
55. Ghosh, P., Krishnamachari, B.: Interference power bound analysis of a network of wireless robots. CoRR (2016). arXiv:1608.08261
56. Dagefu, F.T., Verma, G., Rao, C.R., Paul, L.Y., Fink, J.R., Sadler, B.M.: Sarabandi, Short-range low-vhf channel characterization in cluttered environments. IEEE Trans. Antennas Propag. **63**(6), 2719–2727 (2015)
57. Boillot, N., Dhoutaut, D., Bourgeois, J.: Large scale mems robots cooperative map building based on realistic simulation of nano-wireless communications. Nano Commun. Netw. **6**(2), 51–73 (2015)
58. Qaraqe, K.A., Yarkan, S., Güzelgöz, S., Arslan, H.: Statistical wireless channel propagation characteristics in underground mines at 900 mhz: a comparative analysis with indoor channels. Ad hoc Netw. **11**(4), 1472–1483 (2013)
59. Das, S.R., Belding-Royer, E.M. Perkins, C.E.: Ad hoc on-demand Distance Vector (aodv) Routing (2003)
60. Chakeres, I.D., Belding-Royer, E.M.: Aodv routing protocol implementation design. In: Proceedings of the IEEE International Conference on Distributed Computing Systems Workshops (2004)
61. Deering, S.E.: Internet Protocol, Version 6 (ipv6) Specification (1998)
62. Johnson, D., Hu, Y.-C., Maltz, D.: The dynamic source routing protocol (dsr) for mobile ad hoc networks for ipv4. Technical report, RFC 4728 (2007)
63. Clausen, T., Jacquet, P.: Optimized link state routing protocol (olsr). Technical report, RFC 3626, (2003)
64. Johnson, D., Ntlatlapa, N., Aichele, C.: Simple Pragmatic Approach to Mesh Routing using Batman (2008)
65. Saumitra, M., Das, Y.C., Hu, C.S., Lee, G., Lu, Y.-H.: Mobility-aware ad hoc routing protocols for networking mobile robot teams. J. Commun. Netw. **9**(3), 296–311 (2007)
66. Johnson, D.B., Maltz, D.A.: Dynamic source routing in ad hoc wireless networks. Kluwer Int. Ser. Eng. Comput. Sci. 153–179 (1996)
67. Perkins, C.E., Royer, E.M.: Ad-hoc on-demand distance vector routing. In: Proceedings of the IEEE Workshop on Mobile Computing Systems and Applications (1999)

68. Lee, S.J., Su, W., Gerla, M.: On-demand multicast routing protocol in multihop wireless mobile networks. Mob. Netw. Appl. **7**(6), 441–453 (2002)
69. Abishek, T.K., Chithra, K.R., Ramesh, M.V.: Aer: adaptive energy efficient routing protocol for network of flying robots monitoring over disaster hit area. In: Proceedings of the Wireless and Optical Communications Conference (WOCC) (2012)
70. Matsumoto, A., Asama, H., Ishida, Y., Ozaki, K., Endo, I.: Communication in the autonomous and decentralized robot system actress. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (1990)
71. Tiderko, A., Bachran, T., Hoeller, F., Schulz, D.: Rose-a framework for multicast communication via unreliable networks in multi-robot systems. Robot. Auton. Syst. **56**(12), 1017–1026 (2008)
72. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik **1**(1), 269–271 (1959)
73. Weitzenfeld, A., Martínez-Gómez, L., Francois, J.P., Levin-Pick, A., Obraczka, K., Boice, J.: Multi-robot systems: extending robocup small-size architecture with local vision and ad-hoc networking. In: Proceedings of the IEEE 3rd Latin American Robotics Symposium (LARS) (2006)
74. Liu, H., Nayak, A., Stojmenović, I.: Localized mobility control routing in robotic sensor wireless networks. In: Mobile Ad-Hoc and Sensor Networks, pp. 19–31. Springer (2007)
75. Sugiyama, H., Tsujioka, T., Murata, M.: Qos routing in a multi-robot network system for urban search and rescue. In: Proceedings of the IEEE International Conference on Advanced Information Networking and Applications (2006)
76. Lin, C.R., Liu, J.-S.: Qos routing in ad hoc wireless networks. IEEE J. Sel. Areas Commun. **17**(8), 1426–1438 (1999)
77. Perkins, C.E., Bhagwat, P.: Highly dynamic destination-sequenced distance-vector routing (dsdv) for mobile computers. ACM SIGCOMM. Comput. Commun. Rev. **24**(4), 234–244 (1994)
78. Konolige, K., Ortiz, C., Vincent, R., Agno, A., Eriksen, M., Limketkai, B., Lewis, M., Briesemeister, L., Ruspini, E., Fox, D.: Large-scale robot teams. Multi-Robot Syst. Swarms Intell. Auton. **2**, 193–204 (2003)
79. Bellur, B., Ogier, R.G.: A reliable, efficient topology broadcast protocol for dynamic networks. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) (1999)
80. Ogier, R., Templin, F., Lewis, M.: Topology dissemination based on reverse-path forwarding (tbrpf). Technical report, IETF RFC 3684 (2004)
81. Sliwa, B., Behnke, D., Ide, C., Wietfeld, C.: Bat Mobile: Leveraging Mobility Control Knowledge for Efficient Routing in Mobile Robotic Networks (2016). *arXiv preprint* arXiv:1607.01223
82. Karp, B., Kung, H.-T.: Gpsr: greedy perimeter stateless routing for wireless networks. In: Proceedings of the ACM International conference on Mobile computing and networking (MobiCom) (2000)
83. Son, D., Helmy, A., Krishnamachari, B.: The effect of mobility-induced location errors on geographic routing in mobile ad hoc sensor networks: analysis and improvement using mobility prediction. IEEE Trans. Mob. Comput. **3**(3), 233–245 (2004)
84. Rao, S.A., Pai, M., Boussedjra, M., Mouzna, J.: Gpsr-l: greedy perimeter stateless routing with lifetime for vanets. In: Proceedings of the IEEE International Conference on ITS Telecommunications (ITST) (2008)
85. Mauve, M., Füßler, H., Widmer, J., Lang, T.: Position-based multicast routing for mobile ad-hoc networks. ACM SIGMOBILE Mob. Comput. Commun. Rev. **7**(3), 53–55 (2003)
86. Lochert, C., Hartenstein, H., Tian, J., Fussler, H., Hermann, D., Mauve, M.: A routing strategy for vehicular ad hoc networks in city environments. In: Proceedings of the IEEE Intelligent Vehicles Symposium (2003)
87. Karp, B.: Challenges in geographic routing: sparse networks, obstacles, and traffic provisioning. In: Presentation at the DIMACS Workshop on Pervasive Networking (2001)

88. Mauve, M., Widmer, J., Hartenstein, H.: A survey on position-based routing in mobile ad hoc networks. IEEE Netw. **15**(6), 30–39 (2001)
89. Balasubramanian, A., Levine, B.N., Venkataramani, A.: Replication routing in dtns: a resource allocation approach. IEEE/ACM Trans. Netw. (TON) **18**(2), 596–609 (2010)
90. Balasubramanian, A., Levine, B.N., Venkataramani, A.: Dtn routing as a resource allocation problem. In: ACM SIGCOMM Computer Communication Review, vol. 37, pp. 373–384. ACM (2007)
91. Burgess, J., Gallagher, B., Jensen, D., Levine, B.N.: Maxprop: routing for vehicle-based disruption-tolerant networks. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) (2006)
92. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and wait: an efficient routing scheme for intermittently connected mobile networks. In: Proceedings of the 2005 ACM SIGCOMM Workshop on Delay-tolerant Networking (2005)
93. Spyropoulos, T., Psounis, K., Raghavendra, C.S.: Spray and focus: Efficient mobility-assisted routing for heterogeneous and correlated mobility. In: Proceedings of the IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom) (2007)
94. Li, X., Shu, W., Li, M., Huang, H., Wu, M.-Y.: Dtn routing in vehicular sensor networks. In: Proceedings of the Global Telecommunications Conference (GLOBECOM) (2008)
95. Fall, K., Farrell, S.: Dtn: an architectural retrospective. IEEE J. Sel. Areas Commun. **26**(5), 828–836 (2008)
96. Fall, K.: A delay-tolerant network architecture for challenged internets. In: Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications (2003)
97. Tassiulas, L., Ephremides, A.: Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. IEEE Trans. Autom. Control **37**(12), 1936–1948 (1992)
98. Moeller, S., Sridharan, A., Krishnamachari, B., Gnawali, O.: Routing without routes: the backpressure collection protocol. In: Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks (2010)
99. Dai, J.G., Lin, W.: Asymptotic optimality of maximum pressure policies in stochastic processing networks. Ann. Appl. Probab. **18**(6), 2239–2299 (2008)
100. Shah, D., Wischik, D.: Optimal scheduling algorithms for input-queued switches. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) (2006)
101. Naghshvar, M., Zhuang, H., Javidi, T.: A general class of throughput optimal routing policies in multi-hop wireless networks. IEEE Trans. Inf. Theory **58**(4), 2175–2193 (2012)
102. Banirazi, R., Jonckheere, E., Krishnamachari, B.: Heat-diffusion: pareto optimal dynamic routing for time-varying wireless networks. In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) (2014)
103. Banirazi, R., Jonckheere, E., Krishnamachari, B.: Dirichlet's principle on multiclass multihop wireless networks: minimum cost routing subject to stability. In: Proceedings of the ACM International Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems (2014)
104. Nunez-Martinez, J., Mangues-Bafalluy, J., Portoles-Comeras, M.: Studying practical any-to-any backpressure routing in wi-fi mesh networks from a lyapunov optimization perspective. In: Proceedings of the IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS) (2011)
105. Alresaini, M., Sathiamoorthy, M., Krishnamachari, B., Neely, M.J.: Backpressure with adaptive redundancy (bwar). In: Proceedings of the IEEE International Conference on Computer Communications (INFOCOM) (2012)
106. Nori, S., Deora, S., Krishnamachari, B.: Backip: backpressure routing in ipv6-based wireless sensor networks. USC CENG Technical report CENG-2014-01
107. Ghosh, P., Ren, H., Banirazi, R., Krishnamachari, B., Jonckheere, E.: Empirical evaluation of the heat-diffusion collection protocol for wireless sensor networks. Comput. Netw. **127**, 217–232 (2017)

108. Gnawali, O., Fonseca, R., Jamieson, K., Moss, D., Levis, P.: Collection tree protocol. In: Proceedings of the ACM Conference on Embedded Networked Sensor Systems (2009)
109. Ferrari, F., Zimmerling, M., Thiele, L., Saukh, O.: Efficient network flooding and time synchronization with glossy. In: Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks (2011)
110. Burri, N., Von Rickenbach, P., Wattenhofer, R.: Dozer: ultra-low power data gathering in sensor networks. In: Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks (2007)
111. Ferrari, F., Zimmerling, M., Mottola, L., Thiele, L.: Low-power wireless bus. In: Proceedings of the ACM Conference on Embedded Networked Sensor Systems (2012)
112. Winter, T., et al.: Rpl: Ipv6 Routing Protocol for Low-power and Lossy Networks, Mar 2012
113. Wang, S., Gasparri, A.: Krishnamachari, B: Robotic message ferrying for wireless networks using coarse-grained backpressure control. IEEE Trans. Mob. Comput. **99**, 1–1 (2016)
114. Olfati-Saber, R., Murray, R.M.: Consensus problems in networks of agents with switching topology and time-delays. IEEE Trans. Autom. Control **49**(9), 1520–1533 (2004)
115. Xiao, L., Boyd, S., Lall, S.: A scheme for robust distributed sensor fusion based on average consensus. In: Proceedings of the ACM/IEEE International Conference on Information Processing in Sensor Networks (2005)
116. Franceschelli, M., Gasparri, A.: Gossip-based centroid and common reference frame estimation in multiagent systems. IEEE Trans. Robot. **30**(2), 524–531 (2014)
117. Lin, Z., Francis, B., Maggiore, M.: Necessary and sufficient graphical conditions for formation control of unicycles. IEEE Trans. Autom. Control **50**(1), 121–127 (2005)
118. Yang, P., Freeman, R.A., Lynch, K.M.: Multi-agent coordination by decentralized estimation and control. IEEE Trans. Autom. Control **53**(11), 2480–2496 (2008)
119. Guo, M., Zavlanos, M.M., Dimarogonas, D.V.: Controlling the relative agent motion in multi-agent formation stabilization. IEEE Trans. Autom. Control **59**(3), 820–826 (2014)
120. Fiedler, M.: Algebraic connectivity of graphs. Czechoslovak Math. J. **23**(2), 298–305 (1973)
121. Mohar, B., Alavi, Y.: The laplacian spectrum of graphs. Graph Theory Comb. Appl. **2**, 871–898 (1991)
122. Godsil, C.D., Royle, G., Godsil, C,D.: Algebraic Graph Theory, vol. 207. Springer, New York (2001)
123. Dimarogonas, D.V., Kyriakopoulos, K.J.: Connectedness preserving distributed Swarm aggregation for multiple kinematic robots. IEEE Trans. Robot. **24**(5), 1213–1223 (2008)
124. Notarstefano, G., Savla, K., Bullo, F., Jadbabaie, A.: Maintaining limited-range connectivity among second-order agents. In: Proceedings of the American Control Conference (ACC) (2006)
125. Savla, K., Notarstefano, G., Bullo, F.: Maintaining limited-range connectivity among second-order agents. SIAM J. Control Optim. **48**(1), 187–205 (2009)
126. Spanos, D.P., Murray, R.M.: Robust connectivity of networked vehicles. In: Proceedings of the IEEE Conference on Decision and Control (CDC) (2004)
127. Yao, Z., Gupta, K.: Backbone-based connectivity control for mobile networks. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2009)
128. Gustavi, T., Dimarogonas, D.V., Egerstedt, M., Hu, X.: Sufficient conditions for connectivity maintenance and rendezvous in leader-follower networks. Automatica **46**(1), 133–139 (2010)
129. DeGennaro, M.C., Jadbabaie, A.: Decentralized control of connectivity for multi-agent systems. In: Proceedings of the IEEE Conference on Decision and Control (CDC) (2006)
130. Dimarogonas, D.V., Johansson, K.H.: Bounded control of network connectivity in multi-agent systems. IET Control Theory Appl. **4**(8), 1330–1338 (2010)
131. Sabattini, L., Secchi, C., Chopra, N., Gasparri, A.: Distributed control of multirobot systems with global connectivity maintenance. IEEE Trans. Robot. **29**(5), 1326–1332 (2013)
132. Gasparri, A., Sabattini, L., Ulivi, G.: Bounded control law for global connectivity maintenance in cooperative multirobot systems. IEEE Trans. Robot. **33**(3), 700–717 (2017)
133. Michael, M.: Zavlanos, Alejandro Ribeiro, and George J Pappas. Network integrity in mobile robotic networks. IEEE Trans. Autom. Control **58**(1), 3–18 (2013)

134. Michael, M.: Zavlanos and George J Pappas. Distributed connectivity control of mobile networks. IEEE Trans. Robot. **24**(6), 1416–1428 (2008)
135. Michael, M.: Zavlanos and George J Pappas. Potential fields for maintaining connectivity of mobile networks. IEEE Trans. Robot. **23**(4), 812–816 (2007)
136. Michael, M.: Zavlanos, Magnus B Egerstedt, and George J Pappas. Graph-theoretic connectivity control of mobile robot networks. Proceedings of the IEEE **99**(9), 1525–1540 (2011)
137. Knorn, F., Stanojevic, R., Corless, M., Shorten, R.: A framework for decentralised feedback connectivity control with application to sensor networks. Int. J. Control **82**(11), 2095–2114 (2009)
138. Schuresko, M., Cortés, J.: Distributed motion constraints for algebraic connectivity of robotic networks. J. Intell. Robot. Syst. **56**(1–2), 99–126 (2009)
139. Manfredi, S.: An algorithm for fast rendezvous seeking of wireless networked robotic systems. Ad Hoc Netw. **11**(7), 1942–1950 (2013)
140. Gil, S., Feldman, D., Rus, D.: Communication coverage for independently moving robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2012)
141. Williams, R.K., Gasparri, A., Sukhatme, G.S., Ulivi, G.: Global connectivity control for spatially interacting multi-robot systems with unicycle kinematics. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2015)
142. Mostofi, Y.: Decentralized communication-aware motion planning in mobile networks: An information-gain approach. J. Intell. Robot. Syst. **56**(1–2), 233–256 (2009)
143. Powers, M., Balch, T., et al.: Value-based communication preservation for mobile robots. In: 7th International Symposium on Distributed Autonomous Robotic Systems (2004)
144. Anderson, S.O., Simmons, R., Golberg, D.: Maintaining line of sight communications networks between planetary rovers. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), vol. 3 (2003)
145. Tardioli, D., Mosteo, A.R., Riazuelo, L., Villarroel, J.L., Montano, L.: Enforcing network connectivity in robot team missions. Int. J. Robot. Res. **29**(4), 460–480 (2010)
146. Ning, B., Jin, J., Zheng, J., Law, Y.W.: Connectivity control and performance optimization in wireless robotic networks: Issues, approaches and a new framework. In: Proceedings of the 6th International Conference on Modelling, Identification & Control (ICMIC) (2014)
147. Wang, B.: Coverage Control in Sensor Networks. Springer Science & Business Media (2010)
148. Cortes, J., Martinez, S., Karatas, T., Bullo, F.: Coverage control for mobile sensing networks. IEEE Trans. Robot. Autom. **20**(2), 243–255 (2004)
149. Schwager, M., Julian, B.J., Rus, D.: Optimal coverage for multiple hovering robots with downward facing cameras. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2009)
150. Choset, H.: Coverage for robotics-a survey of recent results. Ann. Math. Artif. Intell. **31**(1), 113–126 (2001)
151. Hazon, N., Kaminka, G.A.: On redundancy, efficiency, and robustness in coverage for multiple robots. Robot. Auton. Syst. **56**(12), 1102–1114 (2008)
152. Kantaros, Y., Zavlanos, M.M.: Communication-aware coverage control for robotic sensor networks. In: Proceedings of the IEEE Conference on Decision and Control (CDC) (2014)
153. Yan, Y., Mostofi, Y.: Communication and path planning strategies of a robotic coverage operation. In: Proceedings of the American Control Conference (ACC) (2013)
154. Ghaffarkhah, A., Mostofi, Y.: Optimal motion and communication for persistent information collection using a mobile robot. In: Proceedings of the IEEE Globecom Workshops (GC Wkshps) (2012)
155. Ghaffarkhah, A., Mostofi, Y.: Path planning for networked robotic surveillance. IEEE Trans. Signal Process. **60**(7), 3560–3575 (2012)
156. Gonzalez-Ruiz, A., Mostofi, Y.: Cooperative robotic structure mapping using wireless measurements-a comparison of random and coordinated sampling patterns. IEEE Sens. J. **13**(7), 2571–2580 (2013)

157. Mostofi, Y.: Cooperative wireless-based obstacle/object mapping and see-through capabilities in robotic networks. IEEE Trans. Mob. Comput. **12**(5), 817–829 (2013)
158. Le Ny, J., Ribeiro, A., Pappas, G.J.: Adaptive communication-constrained deployment of unmanned vehicle systems. IEEE J. Sel. Areas Commun. **30**(5), 923–934 (2012)
159. Williams, R.K., Sukhatme, G.S.: Cooperative multi-agent inference over grid structured markov random fields. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2011)
160. Gasparri, A., Krishnamachari, B., Sukhatme, G.S.: A framework for multi-robot node coverage in sensor networks. Ann. Math. Artif. Intell. **52**(2), 281–305 (2008)
161. Hunkeler, U., Truong, H.L., Stanford-Clark, A.: Mqtt-s—a publish/subscribe protocol for wireless sensor networks. In: Proceedings of IEEE Conference on Communication systems software and middleware and workshops, (COMSWARE) (2008)
162. Jiang, W., Zefran, M.: Coverage control with information aggregation. In: Proceedings of the IEEE Conference on Decision and Control (CDC) (2013)
163. Stump, E., Jadbabaie, A., Kumar, V.: Connectivity management in mobile robot teams. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2008)
164. Tekdas, O., Yang, W., Isler, V.: Robotic routers: algorithms and implementation. Int. J. Robot. Res. **29**(1), 110–126 (2010)
165. De Hoog, J., Cameron, S., Visser, A.: Dynamic team hierarchies in communication-limited multi-robot exploration. In: Proceedings of the IEEE International Workshop on Safety Security and Rescue Robotics (SSRR) (2010)
166. De Hoog, J., Cameron, S., Visser, A.: Selection of rendezvous points for multi-robot exploration in dynamic environments. In: Proceedings of the International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS) (2010)
167. De Hoog, J., Cameron, S., Visser, A.: Role-based autonomous multi-robot exploration. In: International Conference on Advanced Cognitive Technologies and Applications (COGNITIVE) (2009)
168. Yan, Y., Mostofi, Y.: Robotic router formation-a bit error rate approach. In: Proceedings of the Military Communications Conference (MILCOM) (2010)
169. Yan, Y., Mostofi, Y.: Robotic router formation in realistic communication environments. IEEE Trans. Robot. **28**(4), 810–827 (2012)
170. Dixon, C., Frew, E.W.: Maintaining optimal communication chains in robotic sensor networks using mobility control. Mob. Netw. Appl. **14**(3), 281–291 (2009)
171. Goldenberg, D.K., Lin, J., Morse, A.S., Rosen, B.E., Yang, Y.R.: Towards mobility as a network control primitive. In: Proceedings of the ACM International Symposium on Mobile ad hoc networking and computing (MobiHoc) (2004)
172. Tekdas, O., Plonski, P.A., Karnad, N., Isler, V.: Maintaining connectivity in environments with obstacles. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2010)
173. Williams, R.K., Gasparri, A., Krishnamachari, B.: Route Swarm: Wireless network optimization through mobility. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2014)
174. Gil, S., Kumar, S., Katabi, D., Rus, D.: Adaptive communication in multi-robot systems using directionality of signal strength. Int. J. Robot. Res. **34**(7), 946–968 (2015)
175. Tuna, G., Nefzi, B., Conte, G.: Unmanned aerial vehicle-aided communications system for disaster recovery. J. Netw. Comput. Appl. **41**, 27–36 (2014)
176. Fida, A., Iqbal, M., Ngo, T.D.: Communication-and position-aware reconfigurable route optimization in large-scale mobile sensor networks. EURASIP J. Wirel. Commun. Netw. **2014**(1), 1–20 (2014)
177. Kennedy, J.: Particle Swarm optimization. In: Encyclopedia of Machine Learning, pp. 760–766. Springer (2011)
178. Curlander, J.C., McDonough, R.N.: Synthetic Aperture Radar. Wiley, New York, NY, USA (1991)

179. Ghosh, P., Pal, R., Krishnamachari, B.: Towards controllability of wireless network quality using mobile robotic routers. In: CoRR (2016). arXiv:1607.07848
180. Ghosh, P., Krishnamachari, B.: Interference power bound analysis of a network of wireless robots. In: International Conference on Communication Systems and Networks, pp. 7–23. Springer, Cham (2017)
181. Chattopadhyay, A., Coupechoux, M., Kumar, A.: Sequential decision algorithms for measurement-based impromptu deployment of a wireless relay network along a line. IEEE/ACM Trans. Netw. **24**(5), 2954–2968 (2016)
182. Ghosh, A., Chattopadhyay, A., Arora, A., Kumar, A.: As-you-go deployment of a 2-connected wireless relay network for sensor-sink interconnection. In: Proceedings of the International Conference on Signal Processing and Communications (SPCOM) (2014)
183. Zavlanos, M.M., Ribeiro, A., Pappas, G.J.: A framework for integrating mobility and routing in mobile communication networks. In: Conference Record of the Forty Fifth Asilomar Conference on Signals, Systems and Computers (ASILOMAR) (2011)
184. Bezzo, N., Fierro, R., Swingler, A., Ferrari, S.: A disjunctive programming approach for motion planning of mobile router networks. Int. J. Robot. Autom. **26**(1), 13 (2011)
185. Chiu, H.C.H., Shen, W.-M.: Anchor-self-configuring robotic network. In: Proceedings of the IEEE International Conference on Robotics and Biomimetics (ROBIO) (2010)
186. Jung, B., Sukhatme, G.S.: Detecting moving objects using a single camera on a mobile robot in an outdoor environment. In: Proceedings of the International Conference on Intelligent Autonomous Systems (IAS) (2004)
187. Schulz, D., Burgard, W., Fox, D., Cremers, A.B.: Tracking multiple moving targets with a mobile robot using particle filters and statistical data association. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2001)
188. Papanikolopoulos, N.P., Khosla, P.K., Kanade, T.: Visual tracking of a moving target by a camera mounted on a robot: a combination of control and vision. IEEE Trans. Robot. Autom. **9**(1), 14–35 (1993)
189. Lindström, M., Eklundh, J.: Detecting and tracking moving objects from a mobile platform using a laser range scanner. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2001)
190. Marković, I., Chaumette, F., Petrović, I.: Moving object detection, tracking and following using an omnidirectional camera on a mobile robot. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2014)
191. Prassler, E., Scholz, J., Elfes, A.: Tracking multiple moving objects for real-time robot navigation. Auton. Robots **8**(2), 105–116 (2000)
192. Yamano, K., Tanaka, K., Hirayama, M., Kondo, E., Kimuro, Y., Matsumoto, M.: Self-localization of mobile robots with rfid system by using support vector machine. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2004)
193. Hahnel, D., Burgard, W., Fox, D., Fishkin, K., Philipose, M.: Mapping and localization with rfid technology. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2004)
194. Zhou, J., Shi, J.: Rfid localization algorithms and applications-a review. J. Intell. Manuf. **20**(6), 695–707 (2009)
195. Brass, P., Cabrera-Mora, F., Gasparri, A., Xiao, J.: Multirobot tree and graph exploration. IEEE Trans. Robot. **27**(4), 707–717 (2011)
196. Ladd, A.M., Bekris, K.E., Rudys, A., Kavraki, L.E., Wallach, D.S.: Robotics-based location sensing using wireless ethernet. Wirel. Netw. **11**(1–2), 189–204 (2005)
197. Ocana, M.S.J.N.M., Bergasa, L.M., Sotelo, M.A., Nuevo, J., Flores R.: Indoor robot localization system using wifi signal measure and minimizing calibration effort. In: Proceedings of the IEEE International Symposium on Industrial Electronics (2005)
198. Song, D., Kim, C.-Y., Yi, J.: Monte carlo simultaneous localization of multiple unknown transient radio sources using a mobile robot with a directional antenna. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2009)

199. Venkateswaran, S., Isaacs, J.T., Fregene, K., Ratmansky, R., Sadler, B.M., Hespanha, J.P., Madhow, U.: Rf source-seeking by a micro aerial vehicle using rotation-based angle of arrival estimates. In: Proceedings of the American Control Conference (ACC) (2013)
200. Palaniappan, R., Mirowski, P., Ho, T.K., Steck, H., Whiting, P., MacDonald, M.: Autonomous rf surveying robot for indoor localization and tracking. In: International Conference on Indoor Positioning and Indoor Navigation (IPIN) (2011)
201. Dantu, K., Goyal, P., Sukhatme, G.: Relative bearing estimation from commodity radios. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2009)
202. Deyle, T., Nguyen, H., Reynolds, M., Kemp, C.C.: Rf vision: Rfid receive signal strength indicator (rssi) images for sensor fusion and mobile manipulation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2009)
203. Wymeersch, H.: The impact of cooperative localization on achieving higher-level goals. In: Proceedings of the IEEE International Conference on Communications Workshops (ICC) (2013)
204. Koutsonikolas, D., Das, S.M., Hu, Y.C., Lu, Y.-H., George Lee, C.S.: Cocoa: coordinated cooperative localization for mobile multi-robot ad hoc networks. In: Proceedings of the IEEE International Conference on Distributed Computing Systems Workshops (ICDCS) (2006)
205. Zickler, S., Veloso, M.: Rss-based relative localization and tethering for moving robots in unknown environments. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) (2010)
206. Filoramo, C., Gasparri, A., Pascucci, F., Priolo, A., Ulivi, G.: A rssi-based technique for inter-distance computation in multi-robot systems. In: Proceedings of the Mediterranean Conference on Control Automation (MED) (2010)
207. Oliveira, L., Li, H., Almeida, L., Abrudan, T.E.: Rssi-based relative localisation for mobile robots. Ad Hoc Netw. **13**, 321–335 (2014)
208. Vasisht, D., Kumar, S., Katabi, D.: Decimeter-level localization with a single wifi access point. In: Proceedings of the USENIX Symposium on Networked Systems Design and Implementation (NSDI) (2016)
209. VICON. https://www.vicon.com/
210. Gage, D.W.: Network protocols for mobile robot systems. In: Intelligent Systems & Advanced Manufacturing, pp. 107–118. International Society for Optics and Photonics (1998)
211. Fung, C.C., Eren, H., Nakazato, Y.: Position sensing of mobile robots for team operations. In: Proceedings of the IEEE Instrumentation and Measurement Technology Conference (IMTC) (1994)
212. Wilke, P., Bräunl, T.: Flexible wireless communication network for mobile robot agents. Ind. Robot Int. J. **28**(3), 220–232 (2001)
213. Thompson, E.A., McIntosh, C., Isaacs, J., Harmison, E., Sneary, R.: Robot communication link using 802.11 n or 900 MHz OFDM. J. Netw. Comput. Appl. **52**, 37–51 (2015)
214. Fai, Y.C., Amin, S.H.M., Fisal, N., Bakar, J.A.: Bluetooth enabled mobile robot. In: Proceedings of the IEEE International Conference on Industrial Technology (ICIT) (2002)
215. Heidemann, J., Ye, W., Wills, J., Syed, A., Li, Y.: Research challenges and applications for underwater sensor networking. In: Proceedings of the IEEE Wireless Communications and Networking Conference (WCNC), vol. 1, pp. 228–235. IEEE (2006)
216. Ghosh, P., Jagadeesan, N.A., Sakulkar, P., Krishnamachari, B.: Loco: a location based communication scheme. In: MadCom: New Wireless Communication Paradigms for the Internet of Things (2017)
217. Zhang, Q., Cho, W., Sobelman, G.E., Yang, L., Voyles, R.: Twinsnet: a cooperative mimo mobile sensor network. In: Ubiquitous Intelligence and Computing, pp. 508–516. Springer (2006)
218. Hollinger, G.A., Choudhary, S., Qarabaqi, P., Murphy, C., Mitra, U., Sukhatme, G., Stojanovic, M., Singh, H., Hover, F.: Communication protocols for underwater data collection using a robotic sensor network. In: Proceedings of the IEEE GLOBECOM Workshops (GC Wkshps) (2011)

219. Pham, H., Jha, S.: An adaptive mobility-aware mac protocol for sensor networks (ms-mac). In: Proceedings of the IEEE International Conference on Mobile Adhoc and Sensor Systems (MASS) (2004)
220. Ali, M., Suleman, T., Uzmi, Z.A.: Mmac: a mobility-adaptive, collision-free mac protocol for wireless sensor networks. In: Proceedings of the IEEE International Performance, Computing, and Communications Conference (IPCCC) (2005)
221. Jhumka, A., Kulkarni, S.: On the design of mobility-tolerant tdma-based media access control (mac) protocol for mobile sensor networks. In: Distributed Computing and Internet Technology, pp. 42–53. Springer (2007)
222. Zhiyong, T., Dargie, W.: A mobility-aware medium access control protocol for wireless sensor networks. In: Proceedings of the IEEE GLOBECOM Workshops (GC Wkshps) (2010)
223. Gonga, A., Landsiedel, O., Johansson, M.: Mobisense: power-efficient micro-mobility in wireless sensor networks. In: Proceedings of the IEEE International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS) (2011)
224. Nabi, M., Blagojevic, M., Geilen, M., Basten, T., Hendriks, T.: Mcmac: an optimized medium access control protocol for mobile clusters in wireless sensor networks. In: Proceedings of the IEEE Communications Society Conference on Sensor Mesh and Ad Hoc Communications and Networks (SECON) (2010)
225. Dong, Q., Dargie, W.: A survey on mobility and mobility-aware mac protocols in wireless sensor networks. IEEE Commun. Surv. Tutor. **15**(1), 88–100 (2013)
226. Ning, B., Jin, J., Zheng, J., Zhang, H.: Minimizing network interference through mobility control in wireless robotic networks. In: Proceedings of the International Conference on Control Automation Robotics & Vision (ICARCV) (2014)
227. Liu, P.X., Meng, M.Q.-H., Liu, P.R., Yang, S.X.: An end-to-end transmission architecture for the remote control of robots over ip networks. IEEE/ASME Trans. Mechatron. **10**(5), 560–570 (2005)
228. Uchimura, Y., Yakoh, T.: Bilateral robot system on the real-time network structure. IEEE Trans. Ind. Electron. **51**(5), 940–946 (2004)
229. Ping, L., Wenjuan, L., Zengqi, S.: Transport layer protocol reconfiguration for network-based robot control system. In: Proceedings of the IEEE International Conference on Networking, Sensing and Control (2005)
230. Holland, G., Vaidya, N.: Analysis of tcp performance over mobile ad hoc networks. Wirel. Netw. **8**(2/3), 275–288 (2002)
231. Harras, K.A., Almeroth, K.C.: Transport layer issues in delay tolerant mobile networks. In: Networking 2006. Networking Technologies, Services, and Protocols; Performance of Computer and Communication Networks; Mobile and Wireless Communications Systems, pp. 463–475. Springer (2006)
232. Lochert, C., Scheuermann, B., Mauve, M.: A survey on congestion control for mobile ad hoc networks. Wirel. Commun. Mob. Comput. **7**(5), 655–676 (2007)
233. Parker, L.E.: Alliance: an architecture for fault tolerant, cooperative control of heterogeneous mobile robots. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (1994)
234. Volpe, R., Nesnas, I., Estlin, T., Mutz, D., Petras, R., Das, H.: The claraty architecture for robotic autonomy. In: Proceedings of IEEE Aerospace Conference (2001)
235. Arkin, R.C., Balch, T.: Cooperative multiagent robotic systems. Artificial Intelligence and Mobile Robots. MIT/AAAI Press, Cambridge, MA (1998)
236. Stoeter, S.A., Rybski, P.E., Erickson, M.D., Gini, M., Hougen, D.F., Krantz, D.G., Papanikolopoulos, N., Wyman, M.: A robot team for exploration and surveillance: design and architecture. In: Proceedings of the International Conference on Intelligent Autonomous Systems (IAS) (2000)
237. Sanfeliu, A., Andrade-Cetto, J.: Ubiquitous networking robotics in urban settings. In: Proceedings of the Workshop on Network Robot Systems. Toward Intelligent Robotic Systems Integrated with Environments (2006)

238. Sanfeliu, A., Hagita, N., Saffiotti, A.: Network robot systems. Robot. Auton. Syst. **56**(10), 793–797 (2008)
239. Saffiotti, A., Broxvall, M., Gritti, M., LeBlanc, K., Lundh, R., Rashid, J., Seo, B.-S., Cho, Y.-J.: The peis-ecology project: vision and results. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2008)
240. McClure, M., Corbett, D.R., Gage, D.W.L: The darpa landroids program. In: SPIE Defense, Security, and Sensing (2009)
241. Chaimowicz, L., Cowley, A., Gomez-Ibanez, D., Grocholsky, B., Hsieh, M.A., Hsu, H., Keller, J.F., Kumar, V., Swaminathan, R., Taylor, C.J.: Deploying air-ground multi-robot teams in urban environments. In: Multi-Robot Systems. From Swarms to Intelligent Automata Volume III, pp. 223–234. Springer (2005)
242. Dorigo, M., Tuci, E., Groß, R., Trianni, V., Labella, T.H., Nouyan, S., Ampatzis, C., Deneubourg, J.-L., Baldassarre, G., Nolfi, S., et al.: The swarm-bots project. In: Swarm Robotics, pp. 31–44. Springer (2005)
243. Mondada, F., Bonani, M., Guignard, A., Magnenat, S., Studer, C., Floreano, D.: Superlinear physical performances in a swarm-bot. In: Advances in Artificial Life, pp. 282–291. Springer (2005)
244. NECTAR: NEtworked Cooperative Teams of Autonomous Robots. http://webuser.unicas.it/nectar/
245. Seyfried, J., Szymanski, M., Bender, N., Estana, R., Thiel, M., Wörn, H.: The i-swarm project: intelligent small world autonomous robots for micro-manipulation. In: Swarm Robotics, pp. 70–83. Springer. (2005)
246. Woern, H., Szymanski, M., Seyfried, J.: The i-swarm project. In: Proceedings of the IEEE International Symposium on Robot and Human interactive Communication (RO-MAN) (2006)
247. Aquila Project. https://code.facebook.com/posts/268598690180189
248. Project Loon. https://x.company/loon/
249. Swarmrobot. http://www.swarmrobot.org/
250. Swarm-bots. http://www.swarm-bots.org/

Robot and Drone Localization in GPS-Denied Areas

Josh Siva and Christian Poellabauer

Abstract Robots and drones have recently become commonplace, with more advanced and affordable units available every year. While initially they may have been marketed primarily to hobbyists, consumer robots (and drones in particular) have become much more than just toys; they have garnered more and more attention from researchers across fields such as environmental sensing, surveillance, computer vision, machine learning, systems engineering, and networking. Robots and drones provide a rich playground in which to tackle challenging problems aimed at increasing the autonomy of machines. Moreover, as these machines become ever more ubiquitous, there arises both the desire and need to provide a way for them to coordinate their movements and actions so that they can accomplish tasks such as navigating without collision or mapping an area. Such coordination becomes more difficult once the space that must be navigated is in a GPS-denied area. In this chapter, the many facets of robot and drone coordination in GPS-denied areas are discussed, addressing issues associated with localization and coordinating multiple agents as they attempt to accomplish a common goal.

1 Introduction

Consumer and commercial drones and robots have found widespread use and are only going to become more ubiquitous with time. For example, consumer drone sales are expected to increase from 1.9 million in 2016 to 4.3 million in 2020, while commercial drone sales are expected to go from 600,000 to 2.7 million in that same time span [17]. A related trend is the increasing autonomy of cars, which will soon make them more akin to robotic vehicles than traditional transportation systems. These large pools of unmanned systems and vehicles may also be supplemented in

J. Siva (✉) · C. Poellabauer
University of Notre Dame, South Bend, USA
e-mail: jsiva@nd.edu

C. Poellabauer
e-mail: cpoellab@nd.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_17

the future by other mobile robots that will be used to complete a variety of tasks. What we have then is a considerable number of mobile robots, all trying to carry out their tasks in a variety of locations, all the while needing to avoid crashing into obstacles as well as each other.

The coordination of multiple robots and drones is a difficult problem on its own; however, the variety of locations in which they can operate create even more problems with navigation and coordination when such locations lack GPS coverage, such as under dense foliage, in urban areas (urban canyons), and indoors. In these GPS-denied areas, we want robots and drones to be able to, cooperatively or otherwise, continue carrying out their tasks such as search and rescue, navigation, package delivery, or site mapping. This chapter will begin with discussing characteristics of robots and drones (Sect. 2). Since the very essence of multi-robot coordination is cemented in localization, we will explain the more general issue of localization in GPS-denied areas as well as how it pertains to mobile agents and how those agents can act cooperatively (Sects. 5–7). Finally, at the end of the chapter, we will consider localization in the practical context of multi-robot coordination and discuss some of the difficulties associated with coordination in a GPS-denied area (Sect. 8).

2 Robots and Drones

Robots and drones bring to mind many different images ranging from the positronic automatons of science fiction to the remote control quadcopters found in toy stores today. The common viewpoint of a robot or drone is that it is the combination of a computer with some sort of mechanism to allow it to physically manipulate, sense, and move around in its environment. However, once it is also equipped with a form of wireless communication, robots and drones share many similarities and challenges found in sensor networks. Consider that these machines are simply wireless sensor nodes that happen to be mobile, and with this simplified view, it is easy to see that issues such as energy efficiency, communication, message routing, and localization are just as important in networks of robots as they are in networks of sensors, especially when both are deployed in an ad hoc fashion. Another difference that distinguishes robots and drones (or at least the types considered here) from Wireless Sensor Networks is that the former are mission oriented. That is, we are concerned with robots and drones that are trying to accomplish a specific task rather than simply remaining in place and passively collecting data about something. This offers a much richer set of difficulties and solutions when it comes to the overall problem of coordinating groups of drones and robots.[1]

[1]The word drone has historically referred to unmanned aircraft (military aircraft in particular), but aside from the form of propulsion, there is an extraordinary amount of overlap between what we refer to as drones and what we normally call robots. Because of this, as well as the militaristic implications of the word drone, we will hereon refer to flying or airborne robots as simply robots unless the form of propulsion is of importance.

2.1 Autonomy

The degree to which a robot is autonomous is driven by the ability of a robot to carry out the tasks for which it is designed, given the hardware available to it and without human intervention. In particular, we are concerned with achieving a level of navigational autonomy where multiple robots are capable of moving about within a space while avoiding collisions with obstacles and other robots as they try to accomplish various other tasks (though their task could simply be to navigate a space such as when robots are used to develop maps of an area). Note that this chapter is not concerned with robots that exhibit a high level of autonomy when it comes to tasks other than those related to navigation and coordination. For example, a robot that must be told where to go and how to do its job at each step may not be considered autonomous by most people, but if the robot is capable of navigational autonomy, the trip from point A to point B can be handled in an intelligent way that potentially involves coordination with a number of robots. Specifically, *robot coordination* is meant to encompass both collaborative and incidental negotiation of the space by the robots involved. This is the level of autonomy that we wish to examine with respect to movement in GPS-denied areas, and it stands in contrast to non-coordinated robots, which may be individually autonomous, but that do not communicate nor proactively attempt to avoid collision. In the future, robots will likely rely on many techniques to coordinate their actions ranging from sense and avoid to path planning to warning systems [26].

2.2 Form Factors

Terminology aside, the form factor of a robot is important to consider. For example, a nonmobile robot is clearly not concerned with issues of navigation and coordination, and aquatic robots face a drastically different set of difficulties when it comes to that same topic. In particular, we are concerned with coordinating mobile robots on the ground and in the air. Though there are many forms of locomotion that permit robots to be mobile in these domains, it is much easier to consider representative subgroups that illustrate particular mobility features. In particular, ground and air mobility can each be divided into groups of robots that have constrained motion (i.e., planes and cars) and those that have relatively free motion in all directions (i.e., non-fixed wing aircraft and omni-wheel robots). The former group is characterized by potentially higher speeds and more predictable movements, while the latter may move more slowly but can move with more degrees of freedom. These different aspects of motion affect localization/tracking and ultimately how we are going to try to coordinate the movement of these groups of robots.

Beyond the form of mobility, we also expect a certain set of capabilities from robots if we wish them to be able to navigate about a space and coordinate their actions with other robots. For example, there must be a way to both sense and

localize surroundings (walls, chairs, trees, people) as well as other robots. This data must also be communicated in some way to other robots to aid in decision-making and coordination. There are two extremes to the robotic network: first, we can consider a completely centralized system in which the robotic agents have a minimal set of sensors and computational ability and are completely coordinated by a central entity. A strong communication infrastructure is necessary for even the simplest actions as the central entity is responsible for collecting all sensor data (from the robots or elsewhere) and using that to tell the robots what to do. This infrastructure means that the area in which the robots can operate can be restricted and the responsiveness of the robots can be reduced; however, it allows the robots themselves to be extremely simple machines. On the other hand, the second form of robotic network would be completely decentralized and necessitates each robot having sufficient sensors and computational capacity to handle navigation.

Communication in this network then serves to enable the robots to coordinate their actions and even help each other with tasks such as localization. At the extreme, this network would be an ad hoc mesh network. As yet, there is no answer as to the best way robotic networks should be implemented, but there is certainly a precedent in favor of the ad hoc network approach in the form of vehicle to vehicle communication in VANETs. Moreover, a decentralized approach provides an opportunity for resilience by removing the single point of failure of the central entity. On the other hand, it is important to note that there are scenarios in which an implementation lying somewhere in between the two robotic network extremes may be the best solution. Consider the use cases broken out by environment below where the use of multiple robots is useful, the coordination of their movements is crucial, but the ideal network architecture is not always immediately apparent:

- **Office building**—The typical office building provides many obstacles in the form of cubicles, chairs, desks, and people. Best suited to smaller flying robots and those restricted to movement along the ground.
  - Search and Rescue (SaR): Robots respond in a disaster scenario such as a fire or earthquake to quickly search the building for victims, collect information ahead of human first-responders, and provide support for ongoing operations.
  - Mapping and exploration: Whether in the context of providing information to first-responders in SaR or exploring a condemned building, multiple robots can cooperatively plan their investigation of the building to make quick work of the mapping process.
- **Large room**—A single large room such as a warehouse or manufacturing building/room. Airborne robots of larger size are a bit more feasible in this domain.
  - Transportation of goods: Many robots are operating throughout the room, moving pallets of items from one area to another. Their movements are tightly coordinated to prevent collisions while they carry their heavy loads.
- **City**—Tall buildings block Line of Sight (LOS) with GPS satellites leading to either no GPS location updates or locations with terrible accuracy. Robots moving

on the ground need to coordinate their motion with people, cars, and, possibly, other robots. Flying robots of all forms could fill the urban canyons coordinated into lanes or common flight paths.

  - Package delivery/general traffic negotiation: A scenario for the future in which many robots are present in a city and a robot must coordinate its movements with the general robotic and human traffic in order to get where it needs to go (either along roads or in the air) to accomplish its mission.

- **Forest**—A large expanse of trees forming a fairly dense canopy that is occasionally broken up by meadows which may allow for intermittent GPS.

  - SaR: Robots need to scour a forest to find a missing hiker. Their search patterns rely on the fact that the robots know exactly where they are and where they have searched. It is conceivable that some robots are deployed carrying supplies that a lost, hungry, and tired hiker might find useful. These special robots could then be coordinated in a special way by being spread throughout the ranks of other searching robots so that immediate aid is never too far away.

Communication is a very important facet of coordinating multiple robots, but a sufficient treatment of the topic is beyond the scope of this chapter. Other chapters in this book provide an excellent survey of this topic.[2]

3 Localization in GPS-Denied Areas

The topic of localization is the more general issue of locating something within a certain space, meaning that it does not have to be restricted to robots only. One of the most well-known solutions to this particular problem is GPS, where satellites in orbit about the Earth can provide someone with a location within an error margin of a few meters [12]. In general, the goal of localization is to provide an entity with coordinates whether they are absolute as in latitude and longitude or relative to some local frame of reference. However, GPS coverage is not perfect because it is dependent on line of sight with a sufficient number of GPS satellites, which is not guaranteed in places such as dense forests or in urban canyons. When line of sight is not available, then multipath effects can cause errors of over 80 m [12]. The goal then is to find a way to localize a robot without making use of GPS, whether the robot operates almost exclusively in GPS-denied areas or just happened to find itself in one. These two different scenarios imply different requirements for GPS. When a signal is lost, the robot may simply be required to maintain the same localization accuracy as GPS. In contrast, a robot that operates exclusively in GPS-denied areas

[2]The chapter "Robotic Wireless Sensor Networks" dives deeply into various communication challenges and serves as an excellent companion to this chapter.

could very likely find itself in very cluttered, possibly highly dynamic environments such as a warehouse, where the accuracy of localization must be higher than can be achieved from GPS.

4 Technologies for GPS-Less Localization

There are many technologies available that have been researched to replace GPS, either intermittently or completely. By far, the most common approach is to leverage existing wireless networks, but there are many other classes of GPS-less localization, including visual localization, IR-based localization, and sound-based localization to name a few. Of these other classes, visual localization stands out, because of the potential for using the camera(s) for mapping, localization, and collision avoidance. In the case of a flying robot, the payload weight is extraordinarily important. A piece of hardware that can perform multiple tasks simultaneously could keep the payload light and therefore keep the robot in the air for a greater amount of time. This same argument works for radio frequency localization, because it can be used for both localization and communication. It is also important to consider the computational burden that the method of localization places on the robot as this will affect the ability of the robot to handle or react to an event (e.g., a robot trying to enter a flow of traffic) in an appropriate amount of time. Sections 5 and 6 will further explain the details of radio frequency localization and visual localization, respectively, and consider the challenges each approach faces. A final localization method that is often paired with almost any localization strategy is dead reckoning. In dead reckoning, a robot's current position is combined with data from Inertial Measurement Units (IMUs) to estimate the next position. This will be discussed further in Sect. 7.

5 Radio Frequency Localization

Radio frequency localization makes use of wireless communication technologies such as Wi-Fi (IEEE 802.11), Bluetooth (IEEE 802.15.1, Bluetooth SIG), ZigBee (IEEE 802.15.4), or Ultra-Wideband (UWB) (IEEE 802.15.4a) to perform localization. This is particularly advantageous if the communication technology can remain useful as a communication device rather than spending most of its time sending beacons or processing the signal in a way that prevents the data from being captured. Although radio frequency localization is not a particularly new topic, especially in Wireless Sensor Networks, the restriction of only using localization methods that are sufficiently accurate and do not impose a large amount of infrastructure (e.g., many anchors or access points) reduces the number of viable solutions. In the next section, we will discuss the shortcomings of a number of radio frequency localization approaches and show that the prominent candidates for radio frequency localization technologies are Wi-Fi, UWB, and Long-Term Evolution (LTE).

5.1 Problems

Wi-Fi and Bluetooth, including the more current Bluetooth Low Energy (BLE) standard, are convenient tools to use for localization, because they are typically already found in consumer electronics (including consumer robots). Both technologies include a metric called *Received Signal Strength Indicator*, which is an estimate of the strength of the signal as defined by the hardware vendor (at least in the case of 802.11 Wi-Fi). Unfortunately, this also means that its implementation can be somewhat ambiguous [2]. In general, if this value were a completely accurate measurement of the power of the signal at the antenna, the distance to the transmitting device could be easily determined due to the fact that radio signals in free space travel in a predictable way according to

$$P_R = P_0 - 10\gamma log(d) \tag{1}$$

where P_R is the received power in dBm, P_0 is the reference power at 1 m, γ represents various environmental characteristics, and d is the distance between the sender and receiver. However, in realistic settings, the RSSI measurements are not reliable, because they suffer from large errors induced by multipath effects as well as a distinct lack of stability even when the environment is static [19]. Factors such as differences in the radio hardware between manufacturers and the impact of the environment (fading) affect the relationship between distance and signal strength. Due to this high sensitivity to the environment, which can cause localization errors of several meters [41], obtaining accurate localization for the coordination of groups of robots is difficult.

Another approach is to measure the Time of Flight (ToF) of a packet between sender and receiver to estimate the distance. The most important factor to consider for such timing-based approaches is the set of parameters that affect the timing measurement. At a high level, both hardware and software components will introduce delays that must be accounted for. Especially when time stamping is used, accurate synchronization between sender and receiver may also be required. From a statistical point of view, to assess the achievable accuracy of this method, we look to the Cramér-Rao Bound (CRB), which allows us to calculate the estimation variance of a particular method of ranging. The variance on position estimates based on timing is directly proportional to the bandwidth of the communication technology [39, 48], which means that the smaller bandwidths of ZigBee, BLE, and Wi-Fi are, perhaps, not the best choices for timing-based localization. As we will discuss later, this is the reason that UWB communication is a promising technology, because, as its name implies, the bandwidth is large, which enables more precise ToF estimation. Additionally, with channel state information available for Wi-Fi, we will see how revisiting timing-based Wi-Fi localization has paid off (in Sect. 5.2).

Yet another class of solutions frequently found in the literature is fingerprint-based localization, where RSSI values from multiple anchors are used as a fingerprint for a particular position in a space [10, 19, 29]. RSSI measurements are taken at known

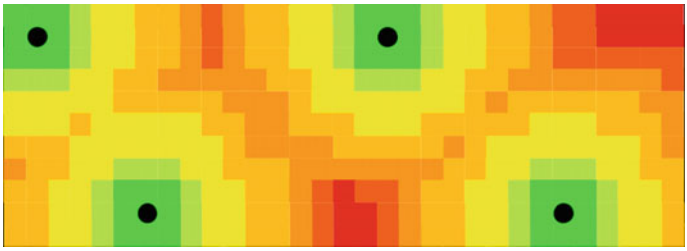


Fig. 1 Example of a radio map used for RSSI fingerprinting

locations and kept in a radio map and RSSI measurements from nodes at unknown locations are compared to these map entries to estimate their locations. An example of such a radio map is shown in Fig. 1, where the black dots indicate the location of a radio transmitter and the changing colors indicate the loss in signal strength with distance to the transmitters. Fingerprint localization can provide good localization accuracy, but it suffers from the same sensitivity to environmental changes that plague other forms of RSSI localization. Although there are solutions to dynamically generate the radio map [29], which can help with the setup cost and the reliability over time, fingerprint-based localization does not offer a great deal of freedom for the devices being localized; that is, they are confined to the mapped area. While it might be okay with a small number of robots restricted to a single room (or highly instrumented building) whose environment is mostly static, this solution does not scale well geographically and simply does not provide the freedom of movement that is often required. Additionally, note that the room must contain a sufficient number of anchors to disambiguate locations on the radio map. This burden of deployment, including setting up and maintaining large numbers of anchors and generating a radio map, is a factor that needs to be seriously considered, especially for a potentially large-scale robot localization system (e.g., large warehouses and urban canyons).

The remaining solutions to radio frequency based localization are all timing based and typically offer excellent accuracy. They are divided into two groups: indoor solutions with Wi-Fi and UWB and an outdoor solution based on LTE. While Wi-Fi- and UWB-based localization could also be used in outdoor environments, LTE-based localization is primarily intended for outdoor use to help replace the loss of GPS in urban canyons. Since LTE signals can penetrate walls of buildings (within reason), it is not out of the question to consider LTE as an indoor localization method; however, from a deployment perspective, it makes much more sense to make use of resources that you have control over. For instance, if Wi-Fi coverage is poor in a particular area, it is possible to move or reconfigure an access point to provide better coverage. The same cannot be said for cell towers. Favoring the more flexible resource for a given area leads us to consider only Wi-Fi and UWB for indoor localization.

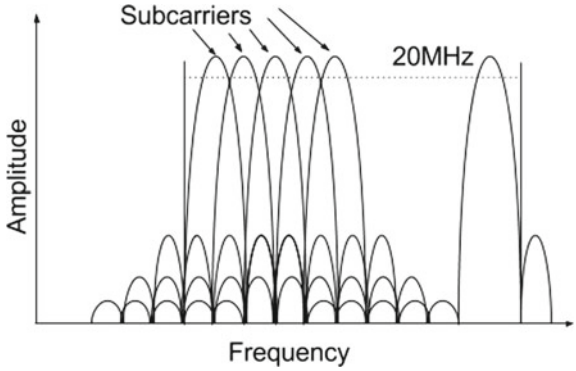


Fig. 2 Subcarriers for OFDM

5.2 Wi-Fi Localization

Starting with 802.11a, Wi-Fi signals were encoded using Orthogonal Frequency Division Multiplexing (OFDM), which uses many subcarriers within a channel to transmit data [16, 23, 49]. A simplified example of this is shown in Fig. 2. For Wi-Fi, each 20 MHz channel consists of 52 subcarriers. A feature found in recent Wi-Fi standards is the use of multiple antennas for spatial diversity and multiplexing. In fact, using multiple antennas, 802.11n introduced the concept of transmit beamforming, where the signal is steered toward the intended target. This was improved upon in 802.11ac [24]. These developments help combat the issue of multipath interference, where reflected signals disrupt the direct path signal at the receiver. As mentioned earlier, multipath propagation is a leading cause of inaccuracies in regular RSSI-based localization schemes, so it would seem that 802.11n would make RSSI localization more appealing. However, it turns out that for transmit beamforming to work, the transmitter must have access to subcarrier level information of the signal. This information is referred to as Channel State Information (CSI). CSI is a collection of physical (PHY) level data that includes complex amplitude and phase as well as time stamps which is a much richer collection of data than RSSI. In recent years, researchers have found ways to access CSI in commercial Wi-Fi Network Interface Cards (NICs), which has opened up a world of possibilities. Some of the NICs that have been used are the Intel 5300 [27], Atheros 9390 [56], and Atheros 9590 [43]. A list of additional Atheros devices can be found at the Atheros CSI Extraction Tool website.[3] The catch with CSI localization is that software to extract it from the chip is not available (at the moment) for anything other than Atheros and the single Intel chipsets. This restriction is a factor to consider when developing a localization scheme for a robotic network as it may be problematic to incorporate the few choices of hardware into a flying robot due to weight, space, and/or hardware compatibility constraints.

[3] http://pdcc.ntu.edu.sg/wands/Atheros/.

With that said, the different approaches to using Wi-Fi CSI for localization can be divided into two main branches: applying traditional RSSI-like path loss localization methods and using methods of localization that, previously, could not be carried out with Wi-Fi. Additionally, the availability of multiple antennas has encouraged researchers in both branches to incorporate methods to differentiate the direct path signal from multipath reflections. For example, in Cupid [56], the Angle of Arrival (AoA) is distinguished by using the MUSIC algorithm [54], which computes the signal intensity over a range of angles.

In line of sight conditions, the greater signal intensity will indicate the AoA of the signal. MUSIC is also used in ToneTrack [67] and SpotFi [35], though not necessarily for the same reason. SpotFi uses a rough Time of Flight (ToF) to determine which signal is the direct path signal and then uses MUSIC to identify the AoA of this signal for localization. Since this ToF value does not take into account the numerous sources of delay, to determine a distance between the sensing node and the target, SpotFi, much like CUPID [40, 56, 66], uses path loss much like in RSSI-based localization except that in this case the signal strength comes from the energy (based on the amplitude of the subcarriers) of the received signals as collected at the PHY level rather than the RSSI value. The median accuracy of these approaches is still on the order of meters, which makes them less appealing for localization in robotic networks operating indoors with the potential for clutter and tight spaces or in a tightly coordinated fashion.

In contrast to the signal strength approaches with CSI above, the new timing-based approaches to Wi-Fi localization are particularly interesting because they have allowed Wi-Fi localization to consistently fall below one meter of error. On the other hand, this is also where we see how big of a difference the choice of hardware makes. For example, SAIL [43], implemented with an Atheros 9590 NIC, demonstrates localization (not just ranging) from a single access point in which AoA is combined with ToF to come up with a location estimate. ToF is calculated based on packet time stamps and provides an estimate with an error of 2 m. ToneTrack [67], implemented with Wireless Open-Access Research Platform (WARP) hardware radios, uses Time of Arrival (ToA) at a set of anchor access points as well as frequency hopping with the MUSIC algorithm to localize a target within less than one meter of error. Chronos [65], implemented with the Intel 5300 NIC, which calculates remarkably accurate ToF to each antenna using a novel approach of aggregating CSI across channels, achieves a ranging error of 10–30 cm between 0 and 15 m and a median localization error between 60 cm and 1m. Finally, Ubicarse [37], implemented with the Intel 5300 NIC on a tablet, combines additional sensor information with CSI obtained from the two antennas on the device to realize a synthetic aperture radar (creating the illusion of an antenna array through movement) with multiple anchors that has a median localization error of 39 cm. While all of these projects use different methods of localization, the large differences in accuracy illustrate how dependent accuracy and reliability can be on the particular chip manufacturer.

Though CSI localization is not without its drawbacks, it has already been successfully incorporated into flying robots [65]. The Atheros Tool was also added to

Manifold Linux which is used by DJI robots.[4] Furthermore, the ability to completely localize a robot with a single anchor (demonstrated in SAIL and Chronos) is particularly useful because it keeps the localization infrastructure minimal, which cannot be said for the localization methods that rely on anchors performing trilateration or triangulation. If the number of hardware options for implementing CSI localization increases, then this will allow greater flexibility for those attempting to build robotic networks with them.

5.3 UWB Localization

UWB communication is characterized by large signal bandwidths (500 MHz and up) and 'bursty' communication that has been designed to provide high throughput as well as a means of localization. The large bandwidth is what makes it possible to use timing-based methods for localization. UWB communication, unlike Wi-Fi and Bluetooth/BLE, is not incorporated into consumer electronics; however, it has found its way into proprietary RF localization solutions such as TimeDomain [5] and Pozyx.[6] More recently, UWB modules have become available for a reasonable price from companies like DecaWave.[7] UWB has been used for decades for communication but has recently garnered a good deal of attention for indoor localization of robots due to its high accuracy. Although UWB devices are still not very common, UWB-based localization systems could easily be used in a robotic network (as they have already been used in single robot applications). Additionally, there is the benefit of avoiding (or working nicely alongside as noted in [3]) the 2.4 GHz band used by Wi-Fi and Bluetooth/BLE, which can help with communication reliability.

The vast majority of UWB localization systems are implemented in an environment where multiple UWB transceivers act as anchors while the robot to be localized carries another one referred to as the tag. Localization transmissions take the form of broadcast from either the anchors or the robot. Though anchor setup does tend to restrict the coverage area for localization, later in Sect. 5.5, we will look at ways for this technology to be extended for multi-agent scenarios.

One of the most simple implementations of this localization environment is presented in [36], where the robot transmits a beacon that is picked up by the anchors. The Time Difference of Arrival (TDOA) at each of the anchors (which are physically wired together to tightly synchronize them) is used to calculate the location of the target. TDOA has an advantage over TOA, because it does not require that the sender and receiver have synchronized clocks. However, this infrastructure creates a bottleneck between the anchors, which could be problematic in terms of scalability, so another approach is to have the robot self-localize. With self-localization, the

[4] https://github.com/libing64/manifold_linux.

[5] http://www.timedomain.com/.

[6] https://www.pozyx.io/.

[7] http://www.decawave.com/.

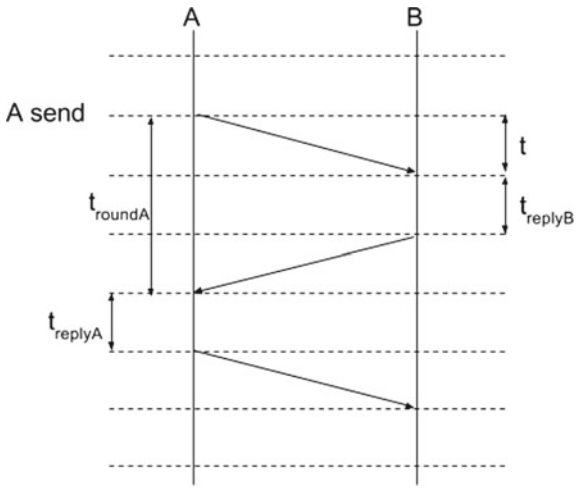


Fig. 3 The message passing sequence for SDS-TWR ranging

anchors transmit a beacon that the robot collects (along with the associated timing information) and then uses TDOA [55] or particle filters [25] to establish its position. Particle filters provide a means of both combining multiple position estimation inputs as well as keeping track of multiple likely candidates for the correct position. Sensor fusion tools such as this are addressed in more depth in Sect. 7.

Another way of tackling localization with UWB is for the robot to use pairwise, rather than broadcast, communication with anchors to establish a range between each anchor and the robot. This technique has been demonstrated in [58, 64], where the method of localization is Symmetric Double Sided Two Way Ranging (SDS-TWR) with multiple anchors surrounding the target (referred to as the "tag"). SDS-TWR refers to the communication that must occur between two points to calculate an accurate Time of Flight that makes up for the lack of synchronization and accounts for clock drift (the changing offset of the clock on the localization target from the anchor time). As illustrated in Fig. 3, given two UWB enabled devices (nodes, robots, etc.), A and B, A sends a message to B, B responds after a certain delay, and A responds to B after a certain delay. These values are combined using the following equation to get the distance:

$$t = t_{roundA} - t_{replyA} + t_{roundB} - t_{replyB}/4$$

Time of Arrival and Two Way Ranging (TWR) are used in [3] and [33], respectively, with less accurate results, which points to SDS-TWR or TDOA as the better methods for robot localization. However, while those two methods provide the best accuracy in terms of localization, it is important to note that they will have detrimental effects on the network performance and scalability. For example, a robot self-localized with SDS-TWR needs to exchange four messages with every anchor (with a minimum of three anchors) to allow it to discover its position. This method

has the potential to cause a bottleneck at the anchors if there are many robots using the same localization method in the same area.

5.4 LTE Localization

LTE signals from cell towers have been proposed as an alternative way to localize aircraft outside when they enter GPS-denied areas. In particular, these areas of interest are urban canyons where the lack of line of site with GPS satellites causes multipath effects that dominate and deteriorate the accuracy. Since cell towers are ubiquitous and provide excellent coverage, they present an interesting opportunity to provide continuous localization outdoors no matter where the robot is located. In addition to radio signals of other sorts such as AM, FM, and Wi-Fi, LTE is considered a Signal Of Opportunity in [45, 57], which can be used to make up for the lack of perfect coverage from GPS.

5.5 Cooperative Radio Frequency Localization

Cooperative localization using radio frequency devices centers around extending absolute localization (localization to a global frame of reference) from areas of coverage (GPS available and/or anchors visible) to areas lacking it and cooperatively correcting position estimates by considering other agents' view of the world. These methods are not mutually exclusive and can help greatly improve the quality of localization in a robotic network. Moreover, some of the methods of cooperative localization have analogs in visual localization.

By using a localization infrastructure (e.g., anchors, GPS, LTE) that cannot reach all robots in the network we create the situation illustrated in Fig. 4. The inner set of robots is not able to localize themselves using the infrastructure, but if the robots labeled A in Fig. 4 are localized then they can act as anchors themselves. This allows all of the robots to localize themselves. Viewed another way, consider a group of robots that have localized themselves relative to each other to produce the (simplified) graph in Fig. 5. This orientation of the robots cannot be properly positioned in the localization space, because all locations are relative to the other robots in the graph. That is, the robots can rotate and translate within the space and there is no measurable difference without an external reference Therefore, the network of robots must be anchored by *absolutely* localizing three of the robots. This would then provide enough information for all of the robots in the network to have absolute positions. For a deeper analysis of this problem from a graph theoretic point of view, see [15].

The secondary use of cooperative localization is to handle noisy measurements (i.e., distance or angle measurements with errors that are large and/or sporadic) so that all robots can be localized. One such example is found in the context of correcting dead reckoning measurements (more on dead reckoning is provided in

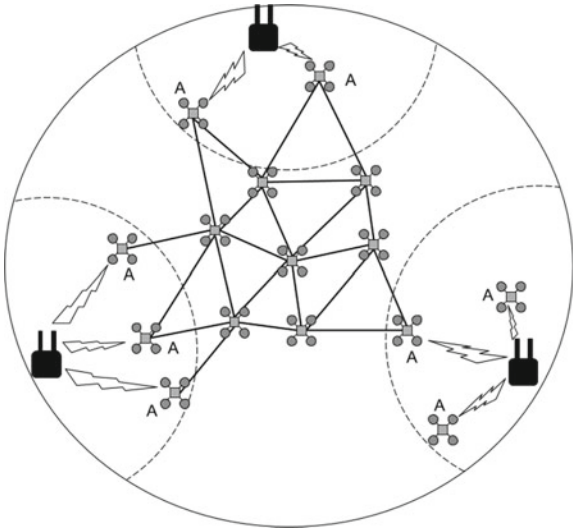


Fig. 4 Robots distributed across an area without GPS except for on the outskirts

Sect. 7), in particular, correcting bearing estimates [31]. In this scenario, the path of a robot is abstracted as a series of lines and joints in which the length of a line (the distance the robot has traveled as calculated through odometry or dead reckoning) is considered fairly trustworthy. Communication distance (or relative localization) is then used between two wandering robots to correct the changes to bearing that occurred at the joints in their paths as illustrated in Fig. 6. In the general case of corrective cooperative localization discussed in [15], we use optimization algorithms to correct distance or bearing measurements according to the constraints imposed by the network. In the case of a network constructed by bearing measurements, this is easily illustrated in Fig. 7 where the bearings pointing to each sensor are corrected so that they intersect in a common location. The usefulness of cooperative localization cannot be overstated as it allows robots to localize themselves even if they are located in a GPS-denied area and it provides a means of correcting localization estimates. Moreover, there is a third aspect of cooperative localization that allows us to construct a better overall picture of a robotic network: the impact on communication. Consider that for all of the robots to be absolutely localized in a GPS-denied area in the absence of cooperative localization, anchors must be provided that cover the entire localization area. This imposes a nontrivial cost to the localization system, because it necessitates either including more anchors or increasing their transmission power. An alternative, then, would be to let a small subset of the robots in the network act as anchors for the rest of the robots. Such a configuration would decrease the number of static anchors that must be introduced to the localization area down to the minimum necessary for absolute localization, but it comes at a cost to the robots assigned as mobile anchors. These robots must have a communication range that covers all of

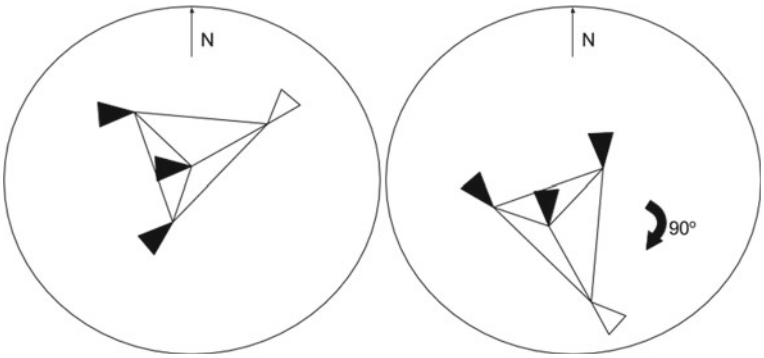


Fig. 5 Localization graph of a set of relatively localized robots

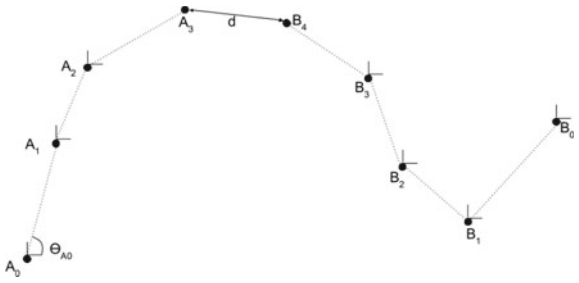


Fig. 6 Position correction by adjusting bearing changes to account for how A met up with B

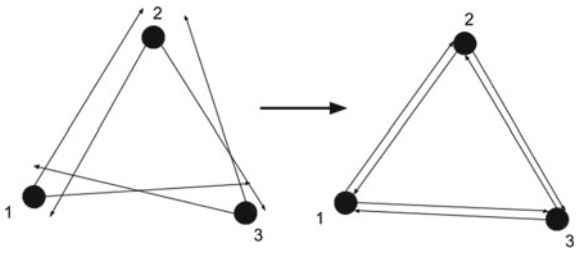


Fig. 7 Corrected set of bearing estimates

the rest of the robots in the network, which would cause an increase in the power consumption of the radio frequency localization devices. Additionally, these robots would face an increased communication/computational burden due to the necessity of carrying out the communication (and maybe computation) required for radio frequency localization. Cooperative localization offers many benefits, but, given the particular deployment scenario, reducing the localization burden on the robots by using a collection of densely deployed anchors could outweigh the cost of setting up such a localization infrastructure. This is a balancing act between scalability and implementation complexity that must be weighed against the needs of the robotic network.

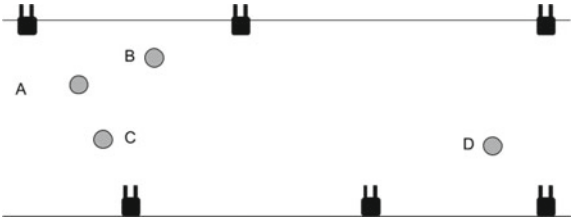


Fig. 8 Infrastructure-based range-free localization in a hallway

5.6 Range-Free Localization

The idea behind range-free localization is that it is possible to use connectivity information that is provided by any sort of wireless communication technology to produce a rough location estimate. These forms of localization are certainly not intended for detecting any sort of fine-grain movement, but they could be useful for quickly determining highly populated/active areas and roughly approximating locations in situations or locations where precise localization would be overkill. Additionally, the assumptions made here are that the communication range greatly exceeds the range where accurate localization is a necessity and simple communication is cheaper (fewer packets transmitted and/or less time spent doing computation) than ranging. Moreover, since ranging/localization accuracy tends to decrease with increased range, the value of a range-based localization estimate may also decrease. That is, a robot is paying the same amount in resources for a less accurate estimate. In this way, range-free localization serves as a way of increasing efficiency in a robotic network.

Range-free localization methods as they would be applied to robotic networks necessitate a certain level of cooperation between the different agents in the network. This cooperation differs from what is discussed in Sect. 5.5, because range-free localization only aims to provide hints about the locations of different robots in the network rather than providing accurate locations. Given an infrastructure-based localization system, we may have the distribution of robots shown in Fig. 8. In this scenario, robots A, B, and C are all located fairly close to one another, so precise localization is certainly desirable to avoid collision. However, robot D is fairly distant so a method of localization such as centroid or APIT [28] would be sufficient. Now if the localization approach is to avoid heavy infrastructure, then we are looking at a scenario that looks very similar to that introduced in Sect. 5.5. Consider the following scenario: Given clusters of robots in a warehouse as shown in Fig. 9, each cluster A, B, C, and D contains robots that are close enough to each other to necessitate accurate localization. However, each cluster pair is beyond this threshold, but they still might like to know where the other clusters are roughly located. How can range-free localization help? One of the most basic methods of range-free localization is called the centroid method which requires (at least) three nodes with known locations that are able to communicate with some new unknown node. The simple way in which this could be used is that clusters A, B, and C all know where they are located relative to

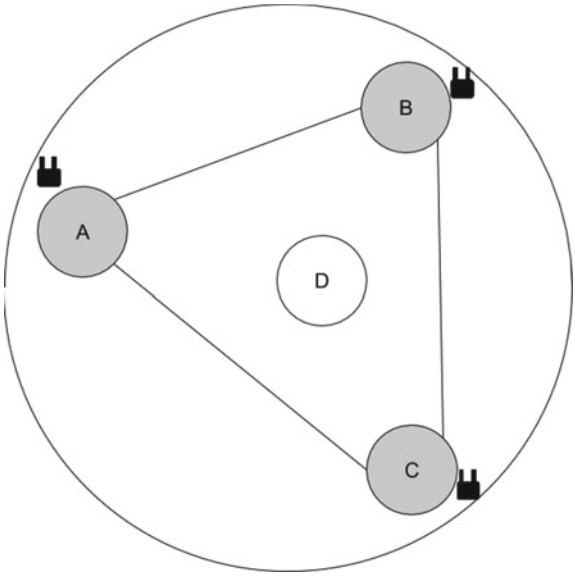


Fig. 9 Rough localization of inner cluster of relatively localized robots

each other (there are beacons/anchors of some sort at these locations) and they guess that cluster D is located at the centroid of the triangle created by A, B, and C. Such an estimate is good enough since, unless D moves toward one of the clusters, it is not important to know exactly where D is located. This example illustrates the biggest drawback to range-free localization: anchors are a necessity. Without agents to act as points of reference there is no way to come up with a position estimate, so range-free localization is purely complementary in a cooperatively localized robotic network; of course, in a densely deployed infrastructure-based robotic network, there are always anchors. In a dense deployment of robots, the centroid method could be improved upon by using APIT for range-free localization, and if we make the assumption that a cooperatively localized robotic network is also a multi-hop network, then we can use other range-free localization algorithms such as DV-Hop [47].

Though range-free localization is based on connectivity information from a form of wireless communication, it does not necessitate the use of radio frequency localization in general. Since any robotic network must have a form of communication, it is not difficult to see how range-free localization could play a complementary role in localization through visual means as well.

As a final note, the performance of the variety of methods presented here is highly dependent upon the technology available/chosen. Moreover, the algorithm, its implementation, and the platform upon which it is implemented further complicate a straight forward comparison between the various approaches. As a starting point in deciding between the different radio localization approaches presented above, Table 1 provides a list of relevant keywords.

Table 1 Summary of RF localization keywords

Ranging	Location	Range-free
TWR	Trilateration	Centroid
SDS-TWR	Triangulation	APIT
MUSIC	Optimization	DV-Hop
RSSI/Power	Filtering	Proximity

6 Visual Localization and Navigation

Visual localization encompasses different, visual, approaches to robot navigation including relative and absolute localization, mapping, and obstacle detection. There are two distinct branches that exist in visual localization: leverage visual markers to assist in the detection of other drones and obstacles or extract characteristics from the images collected by a robot's camera in order to map the environment and assess the robot's motion and location within that map. The first branch is particularly useful if the robot's environment is controlled and it is easy to distribute markers or if markers can easily be applied to other robots. The latter approach allows robots to navigate much more freely and to even generate maps of a location which could be used by other robots (or humans). Such a map, depending on the type, can be invaluable for identifying all of the static (and, to a lesser extent, dynamic) obstacles that exist in a space.

Of course, one of the greatest hurdles to visual localization is that the computational requirements can easily exceed the resources available on a simple robot. To get around this problem, there are four different approaches: offload the computation to an external computer, utilize new technology, reduce the computational burden in software, and increase the processing power available to the robot. Offloading the workload to another computer may be feasible when operating a few robots, but this framework would not scale well; moreover, it is difficult to imagine how it would work with a heterogeneous collection of robots, all with different types of cameras possibly communicating on the same wireless medium. Utilizing new technology such as the Dynamic Visual Sensor from INILabs[8] is certainly attractive due to its ability to provide updates at the microsecond interval and drastically decreased computational requirements when compared to conventional cameras; however, the expense and availability are major concerns at the moment. This technology shows remarkable promise, and will likely play a key role in robotics in years to come. With all this said, we are left with decreasing the computational burden through software and increasing the processing power of the robot, so as we continue to talk about visual localization and navigation, we will simplify the matter by assuming all robots are capable of processing data onboard, and that the key contribution that decreased processing burden provides is a faster update rate leading to greater robot

[8] http://inilabs.com/products/dynamic-vision-sensors/.

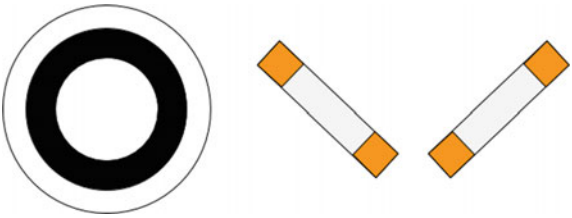


Fig. 10 Examples of custom visual markers

responsiveness. In the rest of this section, we will discuss the major branches of visual localization and navigation as well as methods for visual cooperation.

6.1 Visual Markers

In the land of marker-based visual localization, there are two more subcontinents that divide the topic further. In the first subdivision we have simple, custom markers whose simplicity can allow faster processing of pose estimates (estimates of both the relative location and orientation of a robot) while in the second subdivision, there are fiducial markers which are standardized and can provide information over and above pose.

Custom visual markers come in a variety of shapes and sizes and can be as simple as colored tape placed at particular locations on an obstacle or robot. Examples of custom visual markers are seen in Fig. 10. For a great example of how the simplicity of an image can help with localization computation, consider the image on the left in Fig. 10 which is used in [18]. Through the blob detection algorithm implemented, the robot is able to process a 320×480 stream at 30 Hz on a relatively low powered onboard ARM processor. Additionally, by beginning each iteration of blob detection, in particular, using a method called region growing, starting at the previous center of the blob and stopping when the shape is fully detected, the detection process is sped up greatly. Aside from the ease of detection that this image offers, the shape itself also makes it easy to determine the offset of the camera from the shape. This is accomplished by simply measuring the distortion of the ellipse and the direction in which it occurs to give us a relative location that is within a few centimeters.

Another simple visual marker for relative localization is the one on the right in Fig. 10 [53]. The bright colors at the end of each rectangle make it easy for the robot to detect the marker through image segmentation. The use of a bright color rather than the black and white image above makes the detection of the marker a little bit more robust to lighting conditions. By knowing the distance that the rectangles are separated, the visual distortion caused by the pose of the camera with respect to the marker can be calculated. The combination of location updates paired with onboard sensors through a Kalman Filter (discussed a bit more in Sect. 7) allows the drone to estimate its position within a couple of centimeters.

These two cases illustrate some of the most important factors to consider when designing a custom marker: the difficulty of detection of the shape and color in a variety of environments, the size of the marker, and the camera that will be used for detection. Marker shape and color must be chosen appropriately for the environment in which the marker is going to be used because it is desirable for the marker to be unique. For instance, a rectangular marker may be mistaken for other rectangles which is a very common shape around man-made environments. The size also plays a key role because it, unsurprisingly, impacts the distance at which the marker can be reliably detected, but this, as well as marker shape and color, must also be balanced with the type of camera employed. Cameras differ in their hardware aspects (e.g., monocular vs stereo vs image and depth, frame rate, frame size, lens) and this choice primarily impacts the choice of algorithm, which, in turn affects the processing burden on the robot. A couple informative and concise overviews of camera selection can be found online.[9]

With an understanding of the camera utilized by the robots and how that will impact marker detection, visual localization with custom markers can be rather accurate. However, while custom markers can be quite useful they have a number of drawbacks. For instance, if one is designing a robotic network that includes the use of visual localization, then it becomes necessary to standardize the type of visual marker that will be employed. Additionally, poor image capture or partial occlusion can render the marker undetectable. Both of these problems are (at least partially) solved with the use of fiducial markers; that is, markers that provide some metric against which to judge the captured image. Fiducial markers have proven to be very useful for augmented reality applications because they allow for the resolution of scale in the image as well as camera pose which are useful if a virtual object is to be placed in the scene in a realistic fashion. These two properties also make fiducial markers good for visual localization. A few examples of markers that are used in visual localization are ArUco and AprilTag shown in Fig. 11 [38]. Each of these markers encodes an ID into the bits (the little squares in the images) which can be used for things such as identification or indexing into a table to look up additional information. The encoding of the image using error correction codes provides some robustness to capture quality which is useful for determining pose in uncertain lighting and at different viewing angles. Some fiducial markers are more successful than others in this respect; in particular ArUco and AprilTag markers that are only 3.2 cm^2 are discernible out to 1.2 m with a camera resolution of 640×480 in good lighting conditions whereas the Vuforia marker is only found by the software at a maximum distance of 0.8 m [38]. Moreover, AprilTag is much more resilient to lighting conditions than the other two markers and can even be found at an angle of 60° in all light conditions. It should be noted that Vuforia has gone through a couple iterations since this side by side examination, so the results are not representative of current technology. Still, the point remains that results can vary from one fiducial marker to the next. In addition to using these markers as localization points, these markers can

[9] http://robotsforroboticists.com/camera-lens-selection/, http://www.baslerweb.com/en/vision-campus/camera-selection.

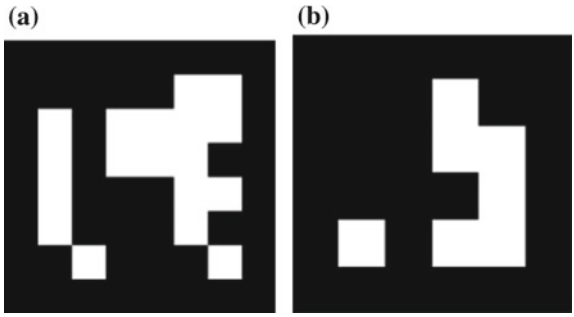


Fig. 11 Examples of fiducials: **a** ArUco **b** AprilTag

be used for identifying obstacles to assist with robot localization [52]. Of course, the drawback to this approach is that all obstacles *must* be identified and tagged with a fiducial marker which is not particularly practical.

Since a visual marker would likely need to be small so as not to be intrusive on the environment (e.g., office, mall, warehouse), the useful area that a robot could navigate in could be prohibitively small. Another place that these markers could be used is on other robots. Such work is demonstrated with custom markers [18, 51] in which one robot is tagged with a visual marker and the other robot calculates the distance between the two. The high precision found in this form of localization allows for more precise movement between multiple robots. The ability to relatively localize in this way is quite similar to relative localization in radio frequency localization except that the direction plays a much bigger role with visual localization.

Notice that all marker-based localization necessitates the robot keeping a marker within its field of view. This means that if tagging robots with markers is the only form of relative localization then there is the potential for there to be a robot in a robotic network that is not observing or being observed by the rest of the network. When this occurs, there is potential for a crash between two robots that did not even know the other was there. On the other hand, tagging all robots with fiducial markers allows for additional information to be conveyed between robots without having to use the potentially busy communication network. Another similarity with RF localization is with the range-free localization methods found in Sect. 5.6. The ability to discern a range from an RF signal is similar to knowing the scale and original shape of a visual marker. Both techniques can lead to accurate positions; however, if less information is known such as only knowing *whether* there is an anchor in range or knowing that a marker exists, then we can turn to range-free localization. That is, being able to recognize markers (because of a distinctive color or shape, perhaps) is enough information to carry out range-free localization. Many of the techniques from Sect. 5.6 will apply for visual markers with some minor alterations to take into account that vision is directional which will likely necessitate taking movement and rotation of the robot into account.

Though the use of visual markers (both custom and fiducials) has its drawbacks with respect to the need to face them and their restricted operating area, the ability to tag places or things with an identifier that allows relative localization could be

very useful for identifying and negotiating an object that must be moved by a robot, identifying and moving about other robots, or perhaps helping a robot to find its way within a warehouse (like a "You are here" symbol on a mall map). Though this approach is not appropriate for all localization purposes in a robotic network, visual markers can certainly serve as an aid to localization. In such a case, visual markers could serve as hints and a tool for measurement correction alongside a more powerful method of localization such as SLAM.

6.2 SLAM

Simultaneous Localization and Mapping (SLAM) is a powerful technique in which a robot explores an area and uses its sensors to develop a map and to simultaneously figure out where the robot is located within this map. While SLAM can be performed both online and offline, we are primarily concerned with online SLAM; that is, performing SLAM while the robot is exploring the area. The overall approach to SLAM is primarily divided between graph theory based methods and filtering methods (filtering will be addressed in the context of dead reckoning in Sect. 7.2 as well). Both approaches attempt to combine the current estimated location along with movement information and data collected about environmental features to solve the SLAM problem. To accomplish this, it is common in SLAM to rely on whatever sensors are used to be able to detect the range and bearing to a (natural or unnatural) landmark in the environment and to be able to uniquely identify the landmark [63]. It is necessary to note that the SLAM problem is not restricted to the use of cameras for the sensing device (e.g., laser range finders are often used), but visual SLAM provides a great deal of information about the explored environment which is why we will focus on it from hereon.

There are two different maps that are generally used for navigation: topological and metric. The former being characterized by its lack of detail and scale so that only important relationships can be seen while the latter is filled with detail (sometimes much more than necessary) and subject to scale so that precise navigation is possible. Though metric maps are the focus for the rest of this section, topological maps certainly have useful characteristics. For instance, topological graphs will take up less storage space than a metric map and may provide sufficient information for path planning on a larger scale where a metric map would be unwieldy. Since the general issue of path planning (rather than tight coordination of robot movement) is not of concern here, we avoid discussing topological maps further though the interested reader can refer to [22].

There are a few different approaches to metric mapping that can be carried out by a robot. In general, the problem reduces to collecting images, producing a 3D position estimate of the pixels in the screen, and recovering the scale of these estimates. As one can imagine, this process is differentiated greatly by whether a monocular, stereo, or depth camera is used. The last two automatically capture depth and scale information, but their useful distance is limited and they are generally more expensive.

While stereo and depth cameras are still worth mentioning, the solutions involving monocular SLAM are particularly interesting due to their minimalist nature (which is ideal for power-constrained robots) so they will be discussed in greater depth here.

First of all, note that the form factor of the robot becomes a little more important when it comes to SLAM techniques. For example, robots that move about on the ground are going to generate a somewhat different map than an airborne robot. Additionally, fixed wing aircraft may find it difficult to perform SLAM for navigating a GPS-denied area due to tighter real-time constraints of a constantly moving aircraft and the potential for a low update frequency. With these considerations in mind, we look at some of the most recent work in monocular SLAM. The most recent advancements in visual SLAM take their inspiration from the work of Klein and Murray [34] who developed Parallel Tracking and Mapping (PTAM). Originally developed for augmented reality, PTAM splits the work of tracking the movement of the camera and developing a map of the space into two threads which run concurrently. To support this, rather than operating on every single frame (which would include a lot of redundant information), the PTAM technique chooses select key frames that are used to optimize the map at a rate that is independent of the frame rate. Two recent works that build on the PTAM framework (as well as work by others in the field) are LSD-SLAM [13] and ORB-SLAM [46]. LSD-SLAM (Large-Scale Direct SLAM) is a method by which optimization is carried out over pixel intensities in the frame rather than using features (special aspects of the scene extracted from a frame) to produce a semi-dense map of the scene. Relying on the whole input image rather than using features extracted from the image allows some robustness to tracking when in either sparsely textured, which makes it difficult to extract features, or highly repetitively textured areas, which leads to the same features being detected again and again; however, LSD-SLAM still uses features for loop detection, that is, figuring when the robot is visiting a location it has already seen. The maps generated by this method are quite accurate and can run in real time on a CPU without the help of a GPU. Moreover, this approach has even been demonstrated on an Android smartphone.[10] Though in LSD-SLAM, the scale of the map is estimated and optimized solely through the accumulation of visual data, SLAM has been implemented for flying robots by one of the same authors using various onboard sensors to help resolve the scale of the map [14]. The other approach that we noted above, ORB-SLAM, relies on features extracted from the frames for tracking, mapping, and loop closure detection. The ORB (Oriented FAST and Rotated BRIEF) feature descriptor is fast to calculate and match; moreover, ORB is also invariant to perspective. ORB-SLAM is demonstrably more accurate than LSD-SLAM and can operate in real-time, and the implementation on a flying robot [42] indicates that it does not require any downsizing of the input frame to maintain a reasonable update rate (15–20Hz) unlike LSD-SLAM. Additionally, with the new ORB-SLAM2 algorithm available,[11] the generation of a semi-dense map like LSD-SLAM is now possible.

[10]For more on LSD-SLAM see http://vision.in.tum.de/research/vslam/lsdslam.

[11]For more information on the ORB-SLAM project visit http://webdiis.unizar.es/~raulmur/orbslam/.

Through the use of powerful SLAM algorithms, robots can both localize themselves and provide a map to aid other robots in localization. Most importantly, the mapping operation only needs to be carried out periodically, so the burden is not on every robot to constantly map the space in which they are operating. Further work still needs to be done to show whether these SLAM methods can be integrated into an autonomous robot. Additionally, it is unclear how these SLAM algorithms will behave during extended missions in a particular area that extend well beyond the discovery of loop closure and the resolution of scale or how robust they are to dynamic environments (though this is touched on briefly in [46]).

6.3 Cooperative Visual Localization

Though it is interesting to see these methods of visual localization for single robots, the real key is to make use of the visual information from a number of robots navigating the same space. This cooperative visual localization can be built upon some of the ideas used above. The first simple example involves the detection of a landmark. Consider two robots A and B moving about in a space such that B is trying to follow A. Now, one way that robot A could assist robot B is to identify landmarks that B should look for to know which direction to go to follow A. There is certainly no reason why a landmark could not be some visual marker we have seen above such as brightly colored tape or a fiducial marker, but there is no reason to restrict ourselves to markers. Robot A could identify a landmark by a set of features such as those used in ORB-SLAM above (the ORB part of ORB-SLAM), but there are other features descriptors such as SIFT, SURF, BRIEF, and FAST. The features can be shared with B (maybe along with the direction that A moved after seeing the landmark), so that B can then search for that landmark and follow along A's path. This is essentially a way of one robot to show another robot what it has seen.

This general idea is taken a step further in multi-robot SLAM in which the visual data is combined across multiple robots to help with map building/optimization including the discovery of loop closures. Using map edges and corners as landmarks, robots in [32] are able to localize themselves relative to a reference map generated by a previous robot's traversal of the area. The odometry drift encountered by the second robot is corrected by matching up the landmarks found by the first robot. Of course, one can make use of more traditional feature descriptors such as SIFT for matching scenes captured by different robots as demonstrated in [50]. In that scenario, the robots are used to implement distributed stereo vision (combining pairs of monocular robots to create the parallax found in stereo vision) and adjust their movement so that they can maximize their overlapping field's of view.

Unfortunately, not all implementations of cooperative SLAM tackle the issue from a distributed approach, which impacts scalability (which may not be an issue depending on the deployment expectations). The Collaborative Structure from Motion (CSfM) system [20] and CoSLAM [70] both process computationally expensive parts of SLAM on an external computer (including GPU acceleration), so it is

difficult to see how such algorithms could be moved to a network of robots with its limited computational capacity. In all fairness, CSfM keeps part of the SLAM computation on the robots themselves (feature extraction and relative motion estimation) while pushing off the computationally expensive fusion of data and mapping to the central computer. Part of the advantage of this is that the robots are not dependent on the central computer for guidance, which gives them some resilience to loss of connectivity or other assorted network issues. On the other hand, there is CoSLAM which is implemented offline with GPU acceleration, but uses the many viewpoints from the robots' cameras to provide excellent resilience to highly dynamic environments. The final approach mentioned here is that of DDF-SAM[12] (Decentralized Data Fusion—Smoothing And Mapping) [8, 9] which takes a strongly statistical approach to fusing the different maps developed on each robot. The method provides robustness to network connectivity issues and operates in a decentralized fashion (as the name implies) which provides hope for greater scalability.

The issue of scalability is very pronounced in cooperative visual localization because visual localization is already computationally demanding. By adding the complications of communication and map merging, the problem can become nearly intractable. Perhaps the one positive to realize in all this is that the computationally demanding task of map optimization and loop closure detection is not a necessity indefinitely, for once an area is explored and mapped it does not need to be continuously reexplored and remapped by the robots that follow afterward. The cost of SLAM, whether cooperative or otherwise, is then more of a one-time setup cost that must be paid for an accurate map of an area. The exploration of an area is not an easy task; a fact to which the human explorers of yore would likely agree.

6.4 Parallels with Visual Sensor Networks

While the assumptions that were made in the beginning of this section allowed us to continue our conversation about visual localization in robots, it is now necessary to step back and consider real-world implementations. Now, if robotic networks are like Wireless Sensor Networks (WSNs), then robotic networks in which cameras are employed are most similar to Visual Sensor Networks (VSNs). These types of networks are characterized by including cameras in the nodes of the WSN and usually performing operations such as occupancy detection, object recognition, and object tracking [59]. Relative visual localization may be used in both robotic networks and VSNs, but, perhaps, the biggest parallel lies in the limitations imposed by power in both. The fact that cannot be overlooked (which is why we keep bringing it up) is that robots do not have an unlimited source of power and not working within the limits can, most especially in the case of flying robots, be catastrophic.

With VSNs, one particularly important issue is how different vision algorithms that run on the nodes are going to impact the battery life of the node. The algorithms

[12] or the more recent DDF-SAM 2.0.

usually employed for computer vision are not lightweight and can take a considerable amount of processing power [1, 6, 59] which means it might be wise to consider when it is *necessary* to utilize such algorithms. A possible way to balance the computation and communication load is to dynamically assign the role of "base station" to an agent within the network. The ad hoc base station can be assigned through a round-robin mechanism or could be determined through market-based, auction-based, or trade-based task assignment methods [68]. One could also consider when it is possible to distribute calculations across multiple nodes rather than trying to stream data constantly to a base station as in CSfM [20]. However, distributing computation means that we must look at the impact of networking and communication on the life of the robot. It is advantageous to minimize, as much as possible, the amount of communication as demonstrated in [44] in which the data association problem (which comes up in the case that landmarks/locations/objects *cannot* be identified perfectly) is distributed across a group of robots. These communication issues share many characteristics not only with VSNs but with MANETs and, to a certain degree, VANETs. Solutions from these similar domains could very well be applied to communication in robotic networks.

7 Dead Reckoning and Filters

7.1 Dead Reckoning

In the absence of constant GPS location updates, localization mechanisms generally fall back on a method of positioning called dead reckoning. In dead reckoning, data from sensors such as accelerometers, gyroscopes, barometers, and ultrasonic range finders are used to figure out where the robot is located relative to its last known location. This allows the robot to estimate its position and attempt to maintain its trajectory even though there are inconsistent location updates. The sensors used in current robots are Micro-Electro-Mechanical Sensors (MEMS) which are packed together into a single unit called an Inertial Measurement Unit (IMU). Dead reckoning turns out to be problematic in the long term because consumer grade IMUs have a tendency to accumulate errors over time. In particular, a key issue with accelerometers is that noisy measurements caused by the necessity of subtracting the acceleration due to gravity are integrated [4] which exacerbates the measurement errors. Moreover, magnetometers are *very* sensitive to electromagnetic field distortions which means that even those created by the robot [7] could be somewhat problematic (though proper placement of hardware on the chassis can alleviate this issue). Electromagnetic interference can come from a variety of sources, which means that the accuracy is often going to be a point of concern.

While dead reckoning on its own, over an extended period of time, may lead to large errors in location estimates, it turns out to be extraordinarily useful in scenarios where location updates are either sporadic, such as when a robot temporarily enters an

urban canyon, or noisy, such as when localization is performed using radio frequency devices indoors. Both sporadic updates and noisy measurements appear in visual localization as well in the form of very low pose estimation update rates (as in SLAM) and when landmarks/markers are not seen clearly by the onboard camera. Dead reckoning can be combined with other localization methods to great effect, especially when used in a way that is referred to as tightly coupled. Tight coupling refers to how the estimates are combined; in particular, tight coupling tends to involve the use of filters such as Kalman Filters (KFs), Extended Kalman Filters (EKFs), or Particle Filters (PF).

7.2 Filtering and Estimation Techniques

The main idea behind KFs, at the highest level, is that they allow one to collect measurements that include noise and make estimates about some property. The estimate is weighted by the level of confidence in a particular measurement. In aiding robotic movement, KFs allow for the estimation of the current position of the robot through Bayesian inference even though the input measurements (such as from the IMU) would produce a very large estimation error on their own. Additionally, KFs can also provide resilience against temporary loss of input, which may be common in, say, a radio frequency localization system. The drawback to KFs is that they are restricted to linear systems which limits their usefulness. To extend KFs to nonlinear systems, that is, systems for which state updates are not all linear equations, derivatives such as EKFs or unscented KFs were developed. However, all variations assume Gaussian error distributions. EKFs in particular have seen a lot of attention in the field of robotics.

A KF or a derivative are not the only choices when it comes to estimating position based on noisy measurements. Another option to tackle systems that exhibit nonlinearity is to use a PF in which the problem of filtering is tackled through a genetic approach. A PF uses a stochastic distribution of particles that is updated based on collected observations. PFs are able to handle almost any probabilistic robot model that can be formulated as a Markov chain which makes them applicable to a much wider range of problems than KFs and its derivatives. The computational complexity is related to the number of particles tracked; therefore, it is possible to decrease the number of particles tracked to allow the algorithm to run on a resource-constrained platform (with decreased accuracy, of course). One of the biggest drawbacks to PFs is that the number of particles increases exponentially with the dimension of the state space [62]; however, it is possible to decrease the complexity of PFs by turning to a family of PFs called Rao-Blackwellized PFs. Such filters use the probabilistic structure of a problem, such as creating conditional independence between feature locations by knowing the path of a robot during SLAM, to greatly decrease the impact of state-space dimension on the computational complexity.

Localization with radio frequency communication benefits greatly from filtering location estimates with dead reckoning (also referred to as fusion). Certainly, noisy

RSSI-based localization is a prime candidate for sensor fusion especially on a smartphone platform [69] where the choices for GPS alternatives are few and far between. Though radio frequency localization is improved tremendously through the use of UWB or CSI data, localization can still benefit from dead reckoning and/or sensor fusion. Consider CUPID [56] and SAIL [43] which both use CSI Wi-Fi localization. Both of these localization methods use dead reckoning (based on accelerometer-derived step counts) to attempt to disambiguate the angle of the direct signal which can not be determined solely from the MUSIC algorithm and the linear array of antennas at the access point. SAIL goes even further and fuses accelerometer, gyroscope, and compass heading data with a KF to estimate the displacement of the user. Using KFs for tracking can also be accomplished with a more minimal set of inputs; that is, the position estimates can be used in conjunction with a KF to provide smoothing and estimates that assume that the target will move in, roughly, the same direction it was moving during the last update. Such a case is illustrated with UWB localization in [64]. Another tactic to take with position estimates is to treat distance measurements in a probabilistic fashion. This is demonstrated in a UWB localization system which uses a PF to estimate the position of mobile robots in both line of sight and non-line of sight scenarios [25].

Visual localization with robots has inevitably lead to the combination of visual estimates of pose with IMU data. Visual marker relative localization and onboard sensor data can be combined quite effectively with a KF to allow a drone to stay within several centimeters of its intended location [53]. However, when a linear estimator will not do, there is the EKF, which has been used fairly often for implementing visual SLAM on robots. Visual SLAM with a monocular camera must estimate seven degrees of freedom in 3D space due to the six possible rigid body transformations in addition to estimating the scale of the map [60]. Assuming that the map is properly scaled through SLAM then it can become a metric map by making use of IMU measurements (notably, altitude measurements) [14]. Using the IMU for metric scale recovery means that it is not necessary to place some object or image of known dimension in the scene, making the system more dynamic. The recovery of the metric scale of the generated map makes SLAM much more useful for humans and robots alike because it allows the map to be related to an absolute coordinate system. Though these filters are often used for sensor fusion, there is no reason why they cannot be used for the estimation of position based solely on one set of measurements. For example, given a set of images of a landscape (essentially a 2D map), it is possible to use a PF to estimate the location of the robot [21]. Finally, filters can also be used for tracking objects of interest; specifically, as demonstrated in [61], distributed cameras can be used to track a soccer ball. Filtering techniques, in addition to dead reckoning, are invaluable tools for improving localization accuracy and also open the door to tracking the movement of robots.

8 Multi-robot Coordination

With this understanding of localization in its many forms (cooperative localization in particular), a concrete example of where it is useful would be quite excellent to explore. Multi-robot coordination is a problem that relies on the existence of accurate localization information to make decisions about how robots should move so that they do not collide with themselves or obstacles in the area. The use cases that we are most concerned with examining are when robots incidentally end up in the same area and when they are intentionally grouped together to accomplish some goal. With both of these use cases, there are two overall approaches that can be taken: centralized coordination and distributed coordination. Centralized coordination is characterized by some infrastructure (quite likely tied into the localization infrastructure) that is responsible for path planning and collision avoidance. On the other hand, distributed coordination relies on the robots themselves to coordinate their movements so as to prevent collisions and allow for the accomplishment of all robots' goals. There are also commonalities between these two coordination structures. For instance, the low-hanging-fruit of traffic management is to partition the area (including airspace) into different lanes in which traffic must fit certain characteristics such as having a similar form factor, max speed, and be moving in the same direction. Assuming that the movements we are interested in coordinating are a bit more chaotic, then another feature that can appear in both forms of coordination is the determination of maximum speed based on the congestion of robots in the space [5]; however, given that we know location estimates can be imperfect, it also makes sense to include that uncertainty into the determination of a robot's maximum speed. These common coordination features aside, below we explore centralized and distributed coordination.

8.1 Centralized Coordination

Whether localization is handled mostly by the robots themselves or by a localization infrastructure, a centralized coordination system can be a good solution especially if the coordination infrastructure can be implemented alongside an existing localization infrastructure or if the burden of deployment is not too great. Furthermore, recalling the two main forms of localization that we considered in Sects. 5 and 6, we know that wireless localization, minimally, requires an anchor to orient the robots' view of their positions to the real world while visual localization can rely on some visual cue to accomplish the same task.

Given the requirement for (at least one) reference anchor in radio frequency localization, it is reasonable to apply centralized coordination in concert with radio frequency localization. However, it is worth noting that there are forms of visual localization that have been implemented using infrastructure. In particular, a system of ceiling mounted cameras could perform localization on tagged robots similar to the

ground truth system in [30], or, as mentioned in Sect. 6.3, a group of robots could send its visual information to a central hub for processing. In all of these cases, there is the possibility of *something* in the localization space that has access to global information about the locations and objectives of all of the robots, which makes coordination from this information sink a very attractive idea especially in scenarios where the localization area is relatively small or restricted to a single (possibly large) room. Specifically, planning of all routes becomes possible once global information is known which is a computationally expensive task [11] best suited for a powerful controlling node. Of course, with a centralized coordinator there is a single point of failure that could be catastrophic for the robotic network. Moreover, there is the potential for a communication bottleneck as information is fed to the sink, which must be handled for the network to be effective.

8.2 Distributed Coordination

Distributed coordination of multiple robots pairs particularly well with cooperative localization in which an aggregation node is not necessary. Additionally, if the robots' operational area is unknown or widespread, distributed coordination is a more fitting means of avoiding collisions. Distributed coordination is not a simple problem and it has a couple prerequisites. First of all, all robots must be able to communicate information to other robots. Second, there must be some common protocol in use for negotiating interactions. Both of these requirements may seem obvious, but they are essential and nontrivial to implement in the real world. The second point, in particular, is worth considering because the means of robot avoidance could be written into a controller in which control laws maintain a specified distance between robots [11]. All robots should be using the same controllers so that there are no conflicts. Moreover, collision avoidance can become even more problematic in GPS-denied areas (indoors, in particular) because these areas tend to be highly dynamic which means that moving obstacles, other than robots, need to be avoided and also communicated to other robots nearby.

The potential for non-line of sight situations between robots created by both dynamic obstacles or walls in a building can be problematic for (cooperative) localization and may create difficulties with wireless communication. Both of these factors directly impact the effectiveness of distributed coordination because of the potential for noisier position estimates and slower updates to nearby robots due to network degradation. Another factor that needs to be kept in mind, and is related to physical area congestion, is the congestion of the network if the communication radius of the robots within a space is excessively large and communication is constantly taking place (e.g., providing periodic trajectory updates) [5]. Ultimately, distributed coordination involves trading implementation complexity for robotic network deployment flexibility.

9 Open Challenges

This chapter only scratches the surface of the challenges and solutions involved in localizing and coordinating multiple robots in GPS-denied areas. An additional issue that needs to be solved is developing a robust communication infrastructure for a highly dynamic and possibly dense network of robots. Communication challenges such as routing, contention management, and quality of service are similar to those faced in VANETs and MANETs, but that does not mean that they are solved. Furthermore, given that robotic networks should not be restricted to a homogeneous collection of robots, the localization and coordination of multiple robots should be able to integrate not only different methods of localization (including mixed radio frequency-visual localization), but different forms of mobility. Specifically, the fact that two sets of robots use two different forms of localization should not be detrimental to the ability to coordinate them. Additionally, if there is a mixed group of robots working in a space, it is possible that there are some that are essentially deaf and blind; that is, there could be robots that do not meet the requirements for communication and/or coordination that are operating within the space as those that do. If this is the case, what are the best policies of handling these intruders? Finally, network and localization security need to be addressed for deployment of robotic networks in the real world.

10 Summary

In this chapter, we have introduced the idea of robots and drones (flying robots) as Mission-Oriented Wireless Sensor Networks. In particular, we addressed the issue of localization of these agents in GPS-denied areas and illustrated solutions in the form of radio frequency localization and visual localization. These localization methods have both been extended to include cooperative forms which are useful for improving localization estimates and providing a means by which distributed coordination can be implemented in a robotic network. Localization estimations can also be improved through the use of filtering techniques and the fusion of position updates through radio frequency or visual means with dead reckoning estimates. Coordinating multiple robots has many challenges and is a massive field in itself, but the cross section of coordination and localization in GPS-denied areas presents a unique set of challenges to be overcome.

References

1. Akyildiz, I.F., Melodia, T., Chowdhury, K.R.: A survey on wireless multimedia sensor networks. Comput. Netw. **51**(4), 921–960 (2007)
2. Bardwell, J.: You believe you understand what you think i said: the truth about 802.11 signal and noise metrics (2004)

3. Bargshady, N., Alsindi, N.A., Pahlavan, K., Ye, Y., Akgul, F.O.: Bounds on performance of hybrid WiFi-UWB cooperative RF localization for robotic applications. In: 2010 IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops), pp. 277–282. IEEE (2010)
4. Bleser, G., Stricker, D.: Advanced tracking through efficient image processing and visual-inertial sensor fusion. Comput. Graph. **33**(1), 59–72 (2009)
5. Bullo, F., Cortes, J., Martinez, S.: Distributed control of robotic networks: a mathematical approach to motion coordination algorithms. Princeton University Press (2009)
6. Charfi, Y., Wakamiya, N., Murata, M.: Challenging issues in visual sensor networks. IEEE Wirel. Commun. **16**(2), 44–49 (2009)
7. Cruz, D., McClintock, J., Perteet, B., Orqueda, O.A., Cao, Y., Fierro, R.: Decentralized cooperative control-a multivehicle platform for research in networked embedded systems. IEEE Control Syst. **27**(3), 58–78 (2007)
8. Cunningham, A., Indelman, V., Dellaert, F.: DDF-SAM 2.0: consistent distributed smoothing and mapping. In: 2013 IEEE International Conference on Robotics and Automation (ICRA), pp. 5220–5227. IEEE (2013)
9. Cunningham, A., Paluri, M., Dellaert, F.: DDF-SAM: fully distributed slam using constrained factor graphs. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 3025–3030. IEEE (2010)
10. Deepesh, P., Rath, R., Tiwary, A., Rao, V.N., Kanakalata, N.: Experiences with using iBeacons for indoor positioning. In: Proceedings of the 9th India Software Engineering Conference, pp. 184–189. ACM (2016)
11. Desai, J.P., Ostrowski, J., Kumar, V.: Controlling formations of multiple mobile robots. In: Proceedings of 1998 IEEE International Conference on Robotics and Automation, vol. 4, pp. 2864–2869. IEEE (1998)
12. Drawil, N.M., Amar, H.M., Basir, O.A.: Gps localization accuracy classification: a context-based approach. IEEE Trans. Intell. Transp. Syst. **14**(1), 262–273 (2013)
13. Engel, J., Schöps, T., Cremers, D. (2014a). LSD-slam: large-scale direct monocular slam. In: European Conference on Computer Vision, pp. 834–849. Springer
14. Engel, J., Sturm, J., Cremers, D.: Scale-aware navigation of a low-cost quadrocopter with a monocular camera. Robot. Auton. Syst. **62**(11), 1646–1656 (2014b)
15. Eren, T.: Cooperative localization in wireless ad hoc and sensor networks using hybrid distance and bearing (angle of arrival) measurements. EURASIP J. Wirel. Commun. Netw. **2011**(1), 1–18 (2011)
16. Ergen, M.: IEEE 802.11 tutorial. University of California Berkeley, 70 (2002)
17. FAA: FAA aerospace forecast fiscal years 2016–2036 (2016). https://www.faa.gov/data_research/aviation/
18. Faigl, J., Krajník, T., Chudoba, J., Přeučil, L., Saska, M.: Low-cost embedded system for relative localization in robotic swarms. In: 2013 IEEE International Conference on Robotics and Automation (ICRA), pp. 993–998. IEEE (2013)
19. Faragher, R., Harle, R.: An analysis of the accuracy of bluetooth low energy for indoor positioning applications. In: Proceedings of the 27th International Technical Meeting of The Satellite Division of the Institute of Navigation (ION GNSS+ 2014), vol. 812, Tampa, FL, USA (2014)
20. Forster, C., Lynen, S., Kneip, L., Scaramuzza, D.: Collaborative monocular slam with multiple micro aerial vehicles. In: 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3962–3970. IEEE (2013)
21. Fox, D., Thrun, S., Burgard, W., Dellaert, F.: Particle filters for mobile robot localization. Sequential Monte Carlo Methods in Practice, pp. 401–428. Springer (2001)
22. Garcia-Fidalgo, E., Ortiz, A.: Vision-based topological mapping and localization methods: a survey. Robot. Auton. Syst. **64**, 1–20 (2015)
23. Gast, M.: 802.11 Wireless Networks: The Definitive Guide. Southeast University Press (2006)
24. Gast, M.S.: 802.11 ac: A Survival Guide. O'Reilly Media, Inc. (2013)
25. González, J., Blanco, J.-L., Galindo, C., Ortiz-de Galisteo, A., Fernandez-Madrigal, J.-A., Moreno, F.A., Martínez, J.L.: Mobile robot localization based on ultra-wide-band ranging: a particle filter approach. Robot. Auton. Syst. **57**(5), 496–507 (2009)

26. Google: Google UAS airspace system overview (2015). https://utm.arc.nasa.gov/documents.shtml
27. Halperin, D., Hu, W., Sheth, A., Wetherall, D.: Tool release: gathering 802.11n traces with channel state information. ACM SIGCOMM. Comput. Commun. Rev. **41**(1), 53–53 (2011)
28. He, T., Huang, C., Blum, B.M., Stankovic, J.A., Abdelzaher, T.: Range-free localization schemes for large scale sensor networks. In: Proceedings of the 9th Annual International Conference on Mobile Computing and Networking, pp. 81–95. ACM (2003)
29. Hossain, A.M., Soh, W.-S.: A survey of calibration-free indoor positioning systems. Comput. Commun. **66**, 1–13 (2015)
30. Howard, A., Mataric, M.J., Sukhatme, G.S.: Cooperative relative localization for mobile robot teams: an ego-centric approach. Technical Report, DTIC Document (2003)
31. Iwase, T., Shibasaki, R.: Infra-free indoor positioning using only smartphone sensors. In: 2013 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–8. IEEE (2013)
32. Jennings, C., Murray, D., Little, J.J.: Cooperative robot localization with vision-based mapping. In: Proceedings of 1999 IEEE International Conference on Robotics and Automation, 1999, vol. 4, pp. 2659–2665. IEEE (1999)
33. Kempke, B., Pannuto, P., Dutta, P.: Polypoint: guiding indoor quadrotors with ultra-wideband localization. In: Proceedings of the 2nd International Workshop on Hot Topics in Wireless, pp. 16–20. ACM
34. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: 6th IEEE and ACM International Symposium on Mixed and Augmented Reality, 2007. ISMAR 2007, pp. 225–234. IEEE
35. Kotaru, M., Joshi, K., Bharadia, D., Katti, S.: SpotFi: decimeter level localization using WiFi. ACM SIGCOMM Comput. Commun. Rev. **45**, 269–282 (2015). ACM
36. Krishnan, S., Sharma, P., Guoping, Z., Woon, O.H.: A UWB based localization system for indoor robot navigation. In: 2007 IEEE International Conference on Ultra-Wideband, pp. 77–82. IEEE (2007)
37. Kumar, S., Gil, S., Katabi, D., Rus, D.: Accurate indoor localization with zero start-up cost. In: Proceedings of the 20th Annual International Conference on Mobile Computing and Networking, pp. 483–494. ACM
38. La Delfa, G.C., Monteleone, S., Catania, V., De Paz, J.F., Bajo, J.: Performance analysis of visualmarkers for indoor navigation systems. Front. Inf. Technol. Electron. Eng. **17**(8), 730–740 (2016)
39. Lanzisera, S., Zats, D., Pister, K.S.: Radio frequency time-of-flight distance measurement for low-cost wireless sensor localization. IEEE Sens. J. **11**(3), 837–845 (2011)
40. Li, Z., Braun, T., Dimitrova, D.C.: A passive WiFi source localization system based on fine-grained power-based trilateration. In: 2015 IEEE 16th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1–9. IEEE (2015)
41. Liu, H., Darabi, H., Banerjee, P., Liu, J.: Survey of wireless indoor positioning techniques and systems. IEEE Trans. Syst. Man Cybern. Part C (Applications and Reviews), **37**(6), 1067–1080 (2007)
42. Loewen, N., Garc, E.O., Mayol-Cuevas, W., et al.: Towards autonomous flight of micro aerial vehicles using ORB-SLAM. In: 2015 Workshop on Research, Education and Development of Unmanned Aerial Systems (RED-UAS), pp. 241–248. IEEE (2015)
43. Mariakakis, A.T., Sen, S., Lee, J., Kim, K.-H.: Sail: single access point-based indoor localization. In: Proceedings of the 12th Annual International Conference on Mobile Systems, Applications, and Services, pp. 315–328. ACM (2014)
44. Montijano, E., Aragues, R., Sagues, C.: Distributed data association in robotic networks with cameras and limited communications. IEEE Trans. Robot. **29**(6), 1408–1423 (2013)
45. Morales, J., Roysdon, P., Kassas, Z.: Signals of opportunity aided inertial navigation. In: Proceedings of ION GNSS Conference (2016)
46. Mur-Artal, R., Montiel, J., Tardós, J.D.: ORB-SLAM: a versatile and accurate monocular slam system. IEEE Trans. Robot. **31**(5), 1147–1163 (2015)

47. Niculescu, D., Nath, B.: Ad hoc positioning system (APS). In: Global Telecommunications Conference, 2001, GLOBECOM'01, IEEE, vol. 5, pp 2926–2931. IEEE (2001)
48. Patwari, N., Ash, J.N., Kyperountas, S., Hero, A.O., Moses, R.L., Correal, N.S.: Locating the nodes: cooperative localization in wireless sensor networks. IEEE Sig. Process. Mag. **22**(4), 54–69 (2005)
49. Phanse, K., Gopinath, K.: A brief tutorial on IEEE 802.11n (2009). http://www.slideshare.net/gopinathkn/80211n-tutorial
50. Piasco, N., Marzat, J., Sanfourche, M.: Collaborative localization and formation flying using distributed stereo-vision. In: IEEE International Conference on Robotics and Automation, Stockholm, Sweden (2016)
51. Rekleitis, I., Babin, P., DePriest, A., Das, S., Falardeau, O., Dugas, O., and Giguere, P.: Experiments in quadrotor formation flying using on-board relative localization (2015)
52. Sanchez-Lopez, J.L., Pestana, J., de la Puente, P., Suarez-Fernandez, R., Campoy, P.: A system for the design and development of vision-based multi-robot quadrotor swarms. In: 2014 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 640–648. IEEE (2014)
53. Santana, L.V., Brandao, A.S., Sarcinelli-Filho, M., Carelli, R.: A trajectory tracking and 3D positioning controller for the AR. drone quadrotor. In: 2014 International Conference on Unmanned Aircraft Systems (ICUAS), pp. 756–767. IEEE (2014)
54. Schmidt, R.: Multiple emitter location and signal parameter estimation. IEEE Trans. Antennas Propag. **34**(3), 276–280 (1986)
55. Segura, M., Hashemi, H., Sisterna, C., Mut, V.: Experimental demonstration of self-localized ultra wideband indoor mobile robot navigation system. In: 2010 International Conference onIndoor Positioning and Indoor Navigation (IPIN), pp. 1–9. IEEE (2010)
56. Sen, S., Lee, J., Kim, K.-H., Congdon, P.: Avoiding multipath to revive inbuilding WiFi localization. In: Proceeding of the 11th Annual International Conference on Mobile Systems, Applications, and Services, pp. 249–262. ACM (2013)
57. Shamaei, K., Khalife, J., Kassas, Z.: Performance characterization of positioning in LTE systems. In: Proceedings of ION GNSS Conference (2016)
58. Silva, B., Pang, Z., Åkerberg, J., Neander, J., Hancke, G.: Experimental study of UWB-based high precision localization for industrial applications. In: 2014 IEEE International Conference on Ultra-WideBand (ICUWB), pp. 280–285. IEEE (2014)
59. Soro, S., Heinzelman, W.: A survey of visual sensor networks. Adv. Multimed. (2009)
60. Strasdat, H., Montiel, J., Davison, A.J.: Scale drift-aware large scale monocular slam. In: Proceedings of Robotics: Science and Systems VI (2010)
61. Stroupe, A.W., Martin, M.C., Balch, T.: Distributed sensor fusion for object position estimation by multi-robot systems. In: Proceedings of 2001 ICRA, IEEE International Conference on Robotics and Automation, 2001, vol. 2, pp. 1092–1098. IEEE (2001)
62. Thrun, S.: Particle filters in robotics. In: Proceedings of the Eighteenth Conference on Uncertainty in Artificial Intelligence, pp. 511–518. Morgan Kaufmann Publishers Inc (2002)
63. Thrun, S., Leonard, J.J.: Simultaneous localization and mapping. Springer Handbook of Robotics. Springer, pp. 871–889 (2008)
64. Tiemann, J., Schweikowski, F., Wietfeld, C.: Design of an UWB indoor-positioning system for UAV navigation in GNSS-denied environments. In: 2015 International Conference on Indoor Positioning and Indoor Navigation (IPIN), pp. 1–7. IEEE (2015)
65. Vasisht, D., Kumar, S., Katabi, D.: Decimeter-level localization with a single WiFi access point. In: 13th USENIX Symposium on Networked Systems Design and Implementation (NSDI 16), pp. 165–178 (2016)
66. Wu, K., Xiao, J., Yi, Y., Chen, D., Luo, X., Ni, L.M.: Csi-based indoor localization. IEEE Trans. Parallel Distrib. Syst. **24**(7), 1300–1309 (2013)
67. Xiong, J., Sundaresan, K., Jamieson, K.: Tonetrack: leveraging frequency-agile radios for time-based indoor wireless localization. In: Proceedings of the 21st Annual International Conference on Mobile Computing and Networking, pp. 537–549. ACM (2015)
68. Yan, Z., Jouandeau, N., Cherif, A.A.: A survey and analysis of multi-robot coordination. Int. J. Adv. Robot. Syst. **10**(12), 399 (2013)

69. Zhuang, Y., El-Sheimy, N.: Tightly-coupled integration of WiFi and MEMS sensors on handheld devices for indoor pedestrian navigation. IEEE Sens. J. **16**(1), 224–234 (2016)
70. Zou, D., Tan, P.: Coslam: collaborative visual slam in dynamic environments. IEEE Trans. Pattern Anal. Mach. Intell. **35**(2), 354–366 (2013)

Middleware for Multi-robot Systems

Yuvraj Sahni, Jiannong Cao and Shan Jiang

Abstract Recent advances in robotics technology have made it viable to assign complex tasks to large numbers of inexpensive robots. The robots as an ensemble form into a multi-robot system (MRS), which can be utilized for many applications where a single robot is not efficient or feasible. MRS can be used for a wide variety of application domains such as military, agriculture, smart home, disaster relief, etc. It offers higher scalability, reliability, and efficiency as compared to single-robot system. However, it is nontrivial to develop and deploy MRS applications due to many challenging issues such as distributed computation, collaboration, coordination, and real-time integration of robotic modules and services. To make the development of multi-robot applications easier, researchers have proposed various middleware architectures to provide programming abstractions that help in managing the complexity and heterogeneity of hardware and applications. With the help of middleware, an application developer can concentrate on the high-level logic of applications instead of worrying about low-level hardware and network details. In this chapter, we survey state of the art in both distributed MRS and middleware being used for developing their applications. We provide a taxonomy that can be used to classify the MRS middleware and analyze existing middleware functionalities and features. Our work will help researchers and developers in the systematic understanding of middleware for MRS and in selecting or developing the appropriate middleware based on the application requirements.

Y. Sahni · J. Cao (✉) · S. Jiang
Department of Computing, The Hong Kong Polytechnic University, Kowloon, Hong Kong
e-mail: csjcao@comp.polyu.edu.hk

Y. Sahni
e-mail: csysahni@comp.polyu.edu.hk

S. Jiang
e-mail: cssjiang@comp.polyu.edu.hk

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_18

1 Introduction

Recent advances in robotics and other related fields have made it feasible for developers to build inexpensive robots. The current trend in robotics community is to use a group of robots to accomplish task objectives instead of using single-robot systems. These group of robots working in collaboration with each other form an ensemble which is commonly referred as a multi-robot system (MRS). Use of MRS provides better scalability, reliability, flexibility, and versatility, and helps in performing any task in a faster and cheaper way as compared to single-robot system [6]. MRS system can be very useful in search and surveillance applications especially for areas which are difficult or impossible for humans to access. Another benefit of MRS is that it has better spatial distribution [97]. Many applications such as underwater and space exploration, disaster relief, rescue missions in hazardous environments, military operations, medical surgeries, agriculture, smart home, etc. can make use of distributed group of robots working in collaboration with each other [6, 50]. It would not only be difficult but may also lead to wastage of resources if such applications are developed using single-robot systems.

The benefits provided by MRS do not come at low cost. MRS is a dynamic and distributed system where different robots are connected to each other using wireless connection. Robots in MRS should collaborate with each other to perform complex tasks such as navigation, planning, distributed computation, etc. but it is not as easy as the systems are usually heterogeneous. Heterogeneity in MRS can arise due to the use of heterogeneous hardware, software, operating system, or communication protocol and standards. Besides, the large number of robots used in the system makes the system development even more complicated. It is extremely difficult for a robotic system developer to develop such complex systems that should be robust, reliable, scalable, and support the real-time integration of heterogeneous components. Developing a complete robotic application requires knowledge from multiple disciplines such as mechanical engineering, electrical engineering, computer science, etc.

These complexities can be reduced by the use of middleware layer. Middleware provides programming abstractions for a developer so that the developer can focus on application logic instead of low-level details [20]. Many middleware architectures have been proposed for MRS. There is a wide range of applications for MRS, and each application has some specific requirements. It is not trivial to develop a middleware for MRS due to peculiar characteristics of MRS and diverse application requirements. The complexity of middleware becomes higher as more features are incorporated. In fact, it is extremely hard to develop a common middleware for all robotic applications [86]. Therefore, it is important to study different types of middleware to help make a better decision while selecting the middleware for an application.

In this chapter, we first study the recent developments in building MRS. We describe the key applications and requirements of MRS. We then describe the design goals and provide a feature tree-based taxonomy of MRS middleware for systematic understanding of middleware. After giving the background of MRS and the motivation for using middleware, we survey state of the art of middleware for MRS.

Although survey of middleware for robotics can be found in literature in [28, 48, 65, 66], they do not focus specifically on middleware for MRS. Besides, many new middleware architectures have been developed and have not been discussed in previous survey papers.

The contributions of this work are as follows:

- We describe the key requirements and applications of MRS. We also show the developments made by robotics community in building distributed MRS. This is useful for researchers and developers who are interested in developing a real testbed for MRS.
- We provide a feature tree-based taxonomy of MRS middleware features. We have considered features corresponding to both middleware and MRS. We utilize the structure of the phylogenetic tree to give a comprehensive framework that can be used by researchers for systematic understanding and comparison of different MRS middlewares. This is the first time such a taxonomy has been given specifically for MRS middleware.
- We have done a comprehensive review of existing middleware for MRS. 14 different middleware examples have been discussed in this work. We have also provided design goals for middleware and analyzed existing works. The review and analysis done in this chapter will be especially useful for beginners who are interested in developing their own multi-robot system. This work can also be used by developers and other researchers in selecting a suitable middleware based on their application requirements.

The remainder of this chapter is as follows. In Sect. 2, we discuss the recent developments in building MRS and provide a classification of robotic applications. In Sect. 3, we discuss the need of middleware for MRS and give some design goals for MRS middleware. In Sect. 4, we provide a feature tree-based taxonomy of MRS middleware. In Sect. 5, we do the comprehensive review of existing middleware for MRS. In Sect. 6, we provide an analysis of existing middleware for MRS.

2 Existing Multi-robot Systems and Applications

This section is divided into two subsections. In Sect. 2.1, we give some key requirements of MRS and then discuss the developments made by the robotic community in building distributed MRS. In Sect. 2.2, we give a classification of robotic applications. We answer two important questions in this section, which are: What is the current stage of development in MRS? and What are the different possible applications of MRS?

(a) Mataric R2 (b) Khepera (c) Khepera II

Fig. 1 Three kinds of robot for multi-robot system in early age

2.1 Existing Multi-robot Systems

Experimental evaluation and validation are important for research in MRS. It often happens that theoretical models and algorithms with perfect simulation results do not work under real-world conditions. In MRS, these divergences are even more amplified compared with single-robot system due to the large number of robots, interactions between robots, and the effects of asynchronous and distributed control, sensing, actuation, and communication. Therefore, it is crucial to build a testbed for MRS to conduct multi-robot research [64]. In this section, we list key requirements of an MRS and show how robotics community has progressed in building distributed MRS over the years.

One of the earliest multi-robot systems is the Mataric R2 robots built in the 1990s (shown in Fig. 1a). They use a group of four robots to demonstrate and verify the group behavior such as foraging, flocking, and cooperative learning [58]. For each Mataric R2 robot, it equips piezoelectric bump sensors for collision detection, two-pronged forklift for picking goods, six infrared sensors for object detection, and radio transceivers for broadcasting up to one byte of data per second. Nearly the same time, the K-Team from Switzerland developed Khepera robot team in 1996 and Khepera II robot team in 1999 [67] shown in Fig. 1b and Fig. 1c, respectively. The size of the robot is reduced from 36-cm long (Mataric R2 robot) to 8-cm long (Khepera and Khepera II). The Khepera II robot has stronger functionality than the Mataric R2 robot such as more powerful computation ability and more reliable wireless communication. Due to the development of electronic technology, the Khepera II robot also has a smaller size.

After the early age, more and more multi-robot systems are built in both laboratory and industry nowadays. Two representative multi-robot systems are Swarmbot [59, 60] developed by McLurkin and iRobot for research purpose in 2004 (shown in Fig. 2a) and Kiva [96] developed by Amazon for warehouse usage in 2007 (shown in Fig. 2b). Also, the research community has organized a lot of multi-robot competitions such as RoboCup for robotic soccer, MAGIC competition for military surveillance,

(a) Swarmbot (b) Kiva

Fig. 2 Two representative multi-robot systems in recent years: Swarmbot for research purpose and Kiva for industry usage

and MicroMouse for maze exploration. A lot of multi-robot systems result from these competitions such as AIBO dog [18], NAO humanoid [33], and Cmdragons [12].

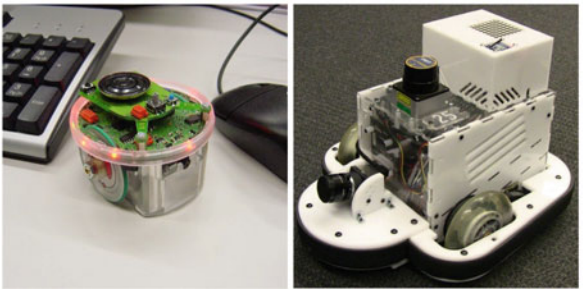
We have observed several features of a multi-robot system:

- Cost: inexpensive for each single robot. A general purpose for MRS is to let quantities of agents, each of which owns limited ability, to achieve a complex system-level target. The system must be designed to be inexpensive to allow researchers to incrementally increase the size of the system. When a multi-robot system is scaled up, it will be hard to cover the fee if each individual robot is highly expensive.
- Size: small size for each single robot. Given limited space, robots with the large size may have problems of frequent collisions, communication blocking, and less flexibility. Also, robots in huge size go against the scalability of the whole system.
- Functionality: stable and strong sensibility for each single robot. If every robot has stable functionality, the whole system can be reliable enough. The stronger the sensibility is, the more the information it may acquire from itself, the environment, and other robots. Hence, the whole system may achieve more complex tasks.

As we know, stronger functionality may result in larger size and higher cost. Therefore, to build an MRS, it is crucial to find a balance between cost, size, and functionality.

Though there have been a lot of multi-robot systems, most of them are controlled in centralized way. In another word, there is a central controller to schedule the robots to perform cooperative tasks. Centralized multi-robot system can be hardly scaled up due to limited computation capability of the central controller. Hence, scholars transfer their research direction to distributed multi-robot system [30]. There are varieties of active research topics that explore efficient algorithms to control distributed multi-robot system, such as self-reconfiguration [7, 76] and exploration [14, 38]. Scholars generally envision their algorithms to be feasible for a distributed multi-robot system consisting of hundreds, thousands, and even more robots

(a) Kilobot (b) R-one (c) Khepera IV



(d) E-puck (e) Scarab

Fig. 3 Five representative robots suitable for multi-robot system nowadays

[7, 25, 90]. However, these algorithms are usually evaluated in simulator only [7, 76], or deployed on a small group of tens of robots or fewer [44, 45] due to cost, time, or complexity. As we previously mentioned, a simulator can hardly model robots' movement, communication, and sensibility in a precise way. Therefore, it would be significant if a large-scale distributed MRS can be built up for algorithm evaluation.

An MRS is said to be fully distributed [72] if each robot in the system supports:

- Distributed control: to process gathered information and to make the decision locally while achieving the system-level goal.
- Distributed sensing: to sense itself, the environment, and other robots locally.
- Distributed actuation: to navigate freely in the environment without collision with obstacles and other robots.
- Distributed communication: to receive and transmit data from other robots in a scalable robot network.

Knowing basic elements for a distributed MRS, we characterize some typical MRSs in detail and compare their functionality and cost. The criteria are to select open-source, still active, and relatively high-impact MRS. The summary of comparison can be seen in Table 1. In detail, five multi-robot systems are considered as follows: Kilobot [77, 79], r-one [61], Khepera IV [88] (evolved from Khepera III [73]), e-puck [68], and Scarab [64].

Table 1 A comparison of off-the-shelf multi-robot systems in terms of functionalities and hardware

	Kilobot	R-one	Khepera IV	E-puck	Scarab
Source	Harvard U	Rice U	K-Team	EPFL	Pennsylvania U
Locomotion	Vibration	Wheel encoders 3-axis gyro 3-axis accelerometer	Wheel encoders 3-axis gyro 3-axis accelerometer	Wheel encoders 3-axis gyro 3-axis accelerometer	Wheel encoders 3-axis gyro 3-axis accelerometer
Sensibility	1 IR range sensor	8 IR range sensors 8 bump sensors 4 light sensors a speaker	8 IR range sensors 8 light sensors 4 IR cliff sensors 5 ultrasonic range sensors 1 microphone 1 speaker 1 color camera	8 IR range sensors 8 light sensors 1 microphone 1 speaker 1 color camera	Laser range sensor high-res color camera
Communication	IR signal	IR signal radio	802.11 b/g Wi-Fi Bluetooth 2.0 EDR	Radio	Radio
Computation	8 MHz Atmega328 32 kB Memory	50 MHz ARM Cortex-M3 64 KB SRAM 256 KB Flash	800 MHz ARM Cortex-A8 512 MB RAM 512 MB flash 4 GB flash for data	Microchip dsPIC MCU 8 KB RAM 144 KB flash	/*
Battery life (h)	3–24	4	7	1–10	/*
Size (cm)	3.3	10	14	7.5	22.2
Cost ($)	14	220	2625	545	3000

*not specified

- The Kilobot[1] (shown in Fig. 3a) is designed by the K-Team and used in SSR lab of Harvard University. Kilobot is a low-cost robotic system especially suitable for research on swarm robotics. The functionality of each individual Kilobot is limited, i.e., only can sense the distance from its neighbor, sense the intensity of visible light, and receive/transfer message from/to its neighbors. However, a collective of Kilobot achieves relatively complicated behaviors such as generating different shapes [80] and transporting large objects [78]. This kind of robotic system in which every robot is with limited ability while can achieve complicated behavior

[1] http://www.eecs.harvard.edu/ssr/projects/progSA/kilobot.html.

together is called swarm robotics. It is inspired by biological swarm behaviors [71] such as bird flocking and ant manipulation. Another such kind of system is the I-Swarm [46] from the University of Stuttgart. However, the robot Jasmine in I-Swarm is far more expensive ($130) compared with Kilobot ($14) while the functionality is similar. Simple functionality makes low cost possible and, on the other hand, limits the feasible environment. For example, a message is transmitted using the reflection of infrared signals. Therefore, the floor where the Kilobots move must be smooth enough, or infrared signals may not reach individual's neighbors.

- The r-one[2] (shown in Fig. 3b) is designed and used in Rice University. The r-one is a relatively low-cost robot that enables large-scale multi-robot research and education. In terms of locomotion, each robot is equipped with two-wheel encoders, a 3-axis gyro, and a 3-axis accelerometer to move on a floor with awareness of odometer, speed, and acceleration. With respect to communication, there are two kinds of communication method. First one is to use infrared transmitter and receiver to achieve directional communication, and the second one is to use radio to achieve nondirectional communication with higher bandwidth. The sensing ability is provided by using 8 bump sensors for 360° detection. r-one provides ample functionalities at a low cost which has motivated its use for education area application [62]. Several courses are taught using r-one. r-one can also be used for multi-robot manipulation [63] and transportation [36] if each robot is equipped with a gripper.
- The Khepera IV[3] (shown in Fig. 3c) is designed and made by K-Team. It is a commercial robot with abundant and powerful functionality compared with non-commercial ones. A standard Khepera IV has the same equipment for locomotion as r-one. For the communication part, Khepera uses 802.11 b/g Wi-Fi and Bluetooth 2.0 EDR for wireless communication instead of infrared signals or radio. Khepera IV has strong sensibility due to the presence of multiple sensors. A Khepera IV is equipped with five ultrasonic transceivers and eight infrared sensors for obstacle detection, four extra infrared sensors for cliff detection, one microphone and one color camera for multimedia functions, and twelve light sensors and three programmable LED for human–robot interaction. Besides, Khepera IV is highly extensible. Developers may extend native functions using the generic USB, Bluetooth devices, and custom boards plugging into the KB-250 bus. Khepera IV wrap the remarkable abilities of sensing, communication, and locomotion in a small body of 14-centimeter diameter. However, the cost of each Khepera IV is over US$ 2600. The Khepera series robot is adopted by DISAL of EPFL and is used for various research topics such as multi-robot learning [26] and odor plume tracing [87].
- The e-puck[4] (shown in Fig. 3d) is designed and made by EPFL. E-puck designer Francesco Mondada started with the Khepera group and moved to make simpler

[2] http://mrsl.rice.edu/projects/r-one.

[3] http://www.k-team.com/khepera-iv.

[4] http://www.e-puck.org/.

education robots. An e-puck is equipped with two-wheel encoders, a VGA camera, three omnidirectional microphones, 3-axis accelerometer, eight infrared sensors, and eight ambient light sensors. Also, e-puck is only 7 cm long and easy to extend functionality. For instance, rotating scanner and turret with three linear cameras are two optional extensions. E-puck is specially designed and widely used for education purpose [21]. It is used in the teaching areas of signal processing, automatic control, behavior-based robotics, distributed intelligent systems, and position estimation and path finding of a mobile robot [68]. In addition, e-puck is also used in many research topics such as supervisory control theory [51] and distributed control strategy [83].

- The Scarab shown in Fig. 3e is designed and made at the University of Pennsylvania. Compared with other robots, the design of Scarab shifts from minimal multi-robots to a complex and robust system. Two of the major components in a Scarab are the Hokuyo URG laser range finder and the Point Grey Firefly IEEE 1394 camera. Using the laser and camera, Scarab is capable of the tasks requiring strong sensibility and high computation payload such as SLAM (simultaneous localization and mapping) [75] and vision processing. However, a Scarab is significantly large, heavy, and expensive with 23 cm diameter, 8 kg weight, and over $3000 cost. Consequently, Scarab is not practical for large populations, i.e., more than ten Scarabs working together. But using less than five Scarabs for multi-robot SLAM is applicable [81].

2.2 Multi-robot System Applications

Robots contain both sensing and actuator components which make them useful for a wide range of applications. Applications which involve navigation, exploration, object transport, and manipulation benefit from the use of MRS. Researchers have been trying to develop biologically inspired robots that incorporate not only the structure of insects and animals but also their social characteristics to design multi-robot system. Researchers try to emulate the communication behavior in bees, birds, and other insects to design control and coordination system for MRS. We have classified the robotic applications into seven categories as shown in Fig. 4. A brief overview of the robotic application is also provided below. These applications are generic and not specifically related to MRS. However, the current research trend is that most applications are now being developed using MRS instead of single-robot system.

- *Healthcare Robots*: Robots have been used by healthcare and medical professionals for a long time. One of the most important uses of robots in health care has been for performing and assisting surgeries. Robots are used for performing precise and minimally invasive surgeries [9, 15]. The current research trend in this area is to use biologically inspired robots that can move in confined spaces and manipulate objects in complex environments [15]. Other areas where robots are being used in

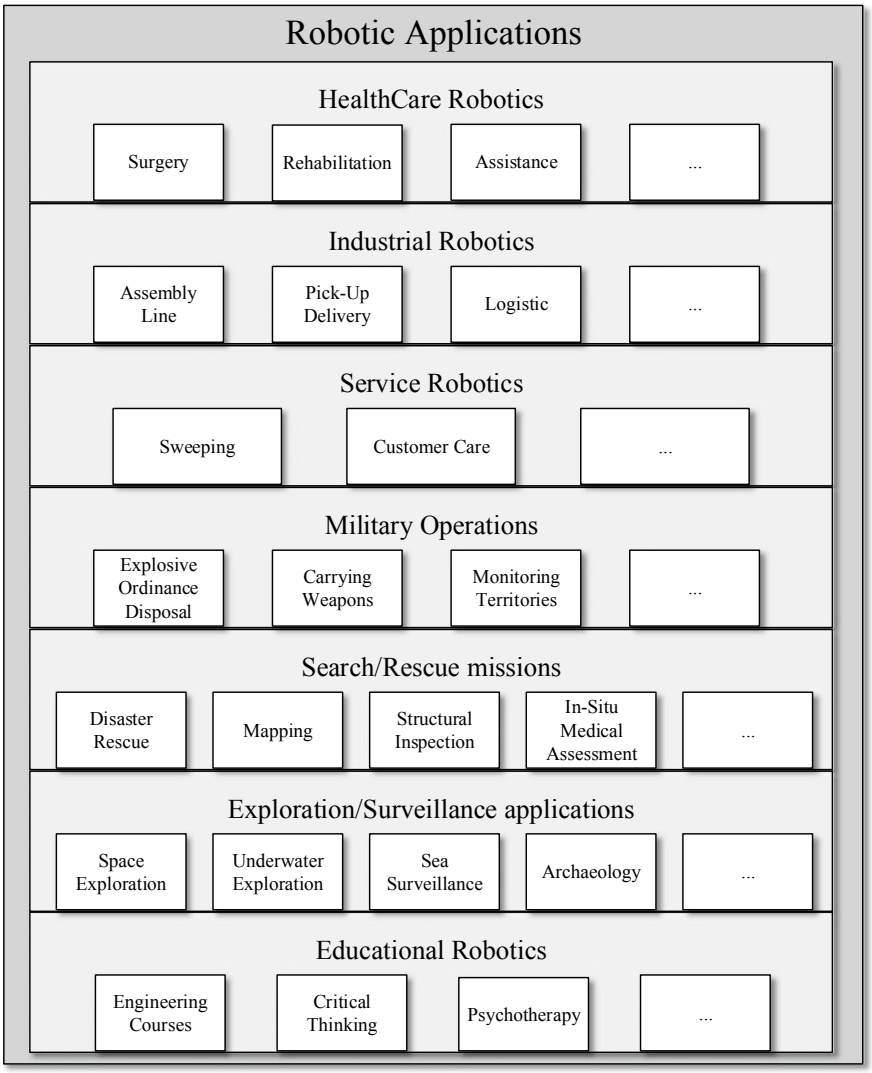


Fig. 4 Classification of robotic application

this application domain is rehabilitation and assistive robotics [34, 91]. Robots are used for recovery of patients with impaired motor and cognitive skills [34]. Robots are being used for assistance to elderly and other physically or mentally disabled individuals to help them live independently. There are even companion robots that help such individuals with special needs. However, due to lack of awareness and other reasons, patients and even healthcare professionals are reluctant to accept robots for medical purpose [11, 91].

- *Industrial robotics*: Robots are now a main component in manufacturing and logistics industries. Industries have been using robots for tasks which are impossible or difficult for humans, such as working in a room filled with hazardous substances, inside a furnace, etc. [34]. Several robotic application studies in manufacturing industries have been mentioned in [29] including die casting applications, forging applications, heat treatment applications, glass manufacturing applications, etc. All large-scale manufacturing industries especially automobile, component assembly, and many other industries involving tasks related to packaging, testing, and logistics rely on the use of robots for efficient task completion [85]. Besides automation, robots are also used for assisting humans in their activities in industries.
- *Service robotics*: Service robots are fully or semi-autonomous robots that perform tasks useful for the well-being of humans except in manufacturing related activities. Service robots are useful for performing tasks that are trivial, dangerous, or repetitive for humans. Home service robots are one such type of robots. They can be used for activities that range from cleaning floor, kitchen, bathroom, windows, swimming pool to lawn mowing, washing clothes, and many other activities [34, 85]. Besides home, service robots can also be used for other services such as object pickup and delivery, customer care, etc. [34].
- *Military operations*: Most of the military organizations around the world are using different types of robots for situations that are risky for humans [82]. Robots are also cheaper to maintain than having the human personnel. Military robots can be classified into three categories, which are ground robots, aerial robots, and maritime robots [82]. These military robots are very often used for battlefield surveillance from ground, air, and underwater level. Ground robots are also being utilized for explosive ordinance disposal. Besides carrying out surveillance operations in enemy territories, unmanned aerial vehicles (UAVs) are also used for carrying missiles to attack enemy sites.
- *Search and Rescue missions*: Rescue robots are used to provide real-time information about the situation to aid search and rescue missions. Rescue robots are used for performing tasks such as searching in unstructured and hazardous environments, reconnaissance and mapping, rubble removal, structural inspection, in-situ medical assessment and intervention, and providing logical support [85]. Rescue robots can be utilized for many situations including natural disasters, mining accidents, fire accidents, explosions, etc. [34]. Rescue robots are also useful for post-disaster experimentation [85]. A key aspect of this application is that rescue robots must be autonomous and they are supposed to work in an unstructured environment where any pre-existing communication network may not work properly.
- *Exploration/Surveillance application*: Robots are useful for collecting data in unstructured environments, unknown territories, and from areas which are difficult or impossible for humans to access. Space exploration, underwater exploration, and exploration in hazardous environments such as radiation prone areas, wilderness, mines, damaged buildings, etc. are some examples of this application [85]. Exploration or surveillance is an important part of other applications too such as military operations, and rescue missions. Navigation, coordination, and

collaboration are three important tasks performed by robots in surveillance applications. A lot of researchers are trying to develop biologically inspired robots that can navigate in confined spaces and perform complex tasks [47].

- *Educational robotics*: Robots are now being used in schools and universities for the educational purpose also. Students can learn about multiple disciplines such as computer science, electronics, mechatronics, etc. by developing robotic applications and learning from the experience [3]. However, there is a drawback with this approach as students only learn about robot-related fields. Several studies have been reviewed in [10], and it is observed that most studies only help in teaching concepts related to physics and mathematics such as Newton's Law of motion, kinematics, fractions, etc. Students who are interested in other fields such as music or arts do not get much benefit out of this. There are few instances where robots have been used for teaching students something different from mathematics or physics. In [95], Lego robots have been used to teach about evolution. Lego robots have also been utilized in [70] to improve social connection in individuals with autism and Asperger's syndrome. This shows that robots have huge potential for contribution toward education. Research efforts are required to find ways to use robots for the development of skills such as critical thinking, problem-solving, teamwork, etc.

3 Design Goals for MRS Middleware

The current trend in robotics is to use MRS for application development instead of a single-robot system. Multiple robots are connected using a wireless network and they work together as a group to accomplish application objectives. These robots are usually composed of heterogeneous hardware and software components that collaborate and coordinate with each other to perform complex tasks such as planning, navigation, distributed computation, object manipulation, etc. [66]. It is not trivial to design software architecture for MRS due to many challenges such as interoperability, dynamic configuration, real-time integration of heterogeneous components, etc. Middleware can resolve these issues by providing programming abstractions and help in reducing the development time and cost [66]. Middleware can also make the application development easier and flexible by providing reusable services. It is, however, challenging to develop a middleware as middleware needs to not only deal with complex issues related to MRS but also satisfy multiple application requirements. In this subsection, we have explained some design goals that should be considered while developing a middleware for MRS. An ideal middleware should be able to support all the features but it should be noted that the complexity of middleware becomes higher as more features are supported. Therefore, it is a trade-off between the number of features supported by a middleware and its complexity.

- *Hardware and software abstractions*: Developing a robotic application requires knowledge of multiple disciplines, which includes knowledge of hardware and

software components being used, and corresponding application domain. Usually, robotic application developers have knowledge of their application domain but it is difficult for them to have expertise on low-level hardware and software issues. The primary purpose of using middleware is to make the application development easier and faster. Development of an application using MRS can be done easily if high-level abstractions are provided to a robotic application developer. Having hardware and software abstractions will enable developers to focus on high-level application requirements rather than low-level hardware and network issues. Besides making the application development easier, it will also help in enhancing the efficiency of the application.

- *Interoperability*: MRSs can have multiple sources of heterogeneity. Heterogeneity in MRS may arise due to the difference in either hardware or software of multiple robots. It is not uncommon to use robots from different hardware manufacturers within the same MRS. Even with the same hardware manufacturer, hardware heterogeneity can arise due to difference in the sensor and actuators being used for the robots. Different communication standards can be used within the same MRS which also leads to heterogeneity in the network. Even if the homogeneous hardware is used for MRS, there can be differences in the software architecture of multiple robots. Software modules developed by different programmers using different programming environments can also lead to heterogeneity. Middleware should provide abstractions for developers to enable interoperability between heterogeneous robots. Middleware should enable platform independence such that robots can be developed on different platforms. Middleware should allow robots developed using different platforms or containing heterogeneous hardware and software components to communicate with each other.
- *Real-time support for required services*: Time-critical robotic applications such as rescue operations, medical surgeries, military operations, etc. require real-time support for services. Most of the applications require real-time support that is required for many services that are responsible for collision detection and avoidance, collaboration between multiple robots in MRS, integration of multiple components in robots, etc. There are some tasks which can afford a delay in services but for most of the services used in MRS, real-time support is required.
- *Dynamic resource discovery and configuration*: MRS is a dynamic system where robots are mobile and since robots are usually used in unstructured environments, there is always change in connectivity. MRS is a scalable system where robots can be added, removed, or changed in configuration. There is always change in the configuration of the network. Middleware should enable dynamic discovery of resources which includes both robots and the software services being used. Middleware should enable autonomous detection and recovery from any fault in the network or software. Middleware should provide support for MRS to be self-adapting, self-configuring, and self-optimizing [66].
- *Flexibility and Software reuse*: Software reuse means using the same service even for a different application, hardware, or software environment. Middleware should enable flexibility in using software services such that services are defined by their functionalities and not based on the hardware, software, or the applications for

which they are used. This implies that a developer should not redevelop the service every time there is some change in hardware, operating system, or even application. Middleware should enable the developer to add new functionalities to a system without having to redevelop everything from scratch.

- *Collaboration among multiple robots*: In MRS, multiple robots collaborate with each other by sharing data. Distributed computation is necessary to enable collaboration between robots; however, due to heterogeneity in MRS, it becomes challenging to understand data belonging to the different types of robots. Another requirement for collaboration between robots is that it should be real-time which makes it even more complex for developers to support this functionality. Middleware should provide services that make it easier to do collaboration between robots. Robots should not only be able to transfer data between each other but also understand the meaning of shared data. Middleware should provide abstractions that can help achieve this objective.
- *Integration with other systems*: Nowadays, robotic applications are developed by integrating robots with other systems such as Internet of things (IoT) and cloud. Cloud robotics is a new paradigm where robots utilize computation and storage benefits of the cloud to perform tasks [39]. In near future, IoT and robotics will be combined to provide better services to humans. Issues and technological implication in implementing IoT-aided robotic applications have been studied in [34]. In coming future, more technologies will be integrated with robotics to provide improved services. Middleware should enable integration of MRS with other systems and technologies. Middleware should provide abstractions for a developer to integrate different technologies.
- *Management and monitoring tools*: A lot of components are involved in development, deployment, and functioning of MRS including multiple robots consisting sensors and actuators, software services, and many other resources. Due to the complexity of MRS, it is difficult for a developer to control everything unless there are some tools available that can help in management and monitoring of the overall system. Besides providing services to program MRS, middleware can also provide management tools to configure, debug, and view the overall MRS [20]. Middleware should also enable the developer to view whole system component-wise to provide a better understanding. This functionality will make it easier even for non-programmers to understand and contribute to the development of the robotic application.
- *Support for the addition of extra services*: MRS is usually deployed in an unstructured environment and every application requires some specific services. Middleware should be flexible to enable the addition of services at runtime. Middleware should support the addition of new services to address both network-specific and application-specific qualities of service (QoS) requirements. It should support the addition of services to address issues such as security, reliability, availability, energy optimization, collision detection and avoidance, etc.

4 A Taxonomy of MRS Middleware

There are tens of existing middleware for multi-robot systems focusing on various aspects and purposes. Among the off-the-shelf middleware, it is difficult for a beginner to choose an appropriate one suitable for a specific multi-robot system or multi-robot application. To address this issue, we propose a taxonomy of MRS middleware features to formally describe MRS middleware. In detail, we utilize the structure of the phylogenetic tree to provide a comprehensive, yet succinct framework that allows for a systematic comparison of MRS middleware. Developers could look up desired features in the phylogenetic tree for the purpose of finding a suitable MRS middleware. In biology, a phylogenetic tree or evolutionary tree is a branching diagram showing the inferred evolutionary relationships among various biological species [24]. In the field of computer science, phylogenetic tree has been used to visualize a taxonomy in many survey papers such as WSN programming abstractions [69], WSN middleware [94], and programming distributed Intelligent MEMS [49], but it has not been used for describing MRS middleware yet.

In Fig. 5, we decompose the MRS middleware features into ten leaf features. Between Fig. 6 and Fig. 13, we describe each leaf feature appeared in Fig. 5 in detail. In these figures, we utilize some notations to describe relationship among features. The relationship between a father feature and several child features can be either inclusive or alternative, notated by solid dot and hollow dot, respectively. Also, a child feature can be either necessary or optional, notated by solid square and hollow square, respectively.

As shown in Fig. 5, when we are investigating MRS middleware features, it can be divided into software features from middleware and hardware features from MRS. On the one hand, features from middleware can be divided into two parts. One is the services by the middleware and the other is the system architecture of the middleware. In terms of provided services, it includes functional services as well as nonfunctional services. With respect to features from the system architecture, three parts are included which are programming abstraction features, infrastructure features, and coordination method features, respectively. On the other hand, features from MRS come from both the infrastructure and concrete applications. The features from infrastructure can be node-level one and system-level one. The features from concrete applications are divided into subcategories based on environment, scope/area, and purpose/goal. In this way, MRS middleware features are divided level by level and result in ten leaf features. We explain and describe each leaf feature in detail in the following paragraphs.

There are a variety of functional features (shown in Fig. 6) for MRS middleware. Functional features of MRS middleware are the basic functions implemented by the middleware. Such functions include localization, mapping, collision avoidance, path planning, vision processing, and many others. With the off-the-shelf implementation, developers may use these basics but important functions conveniently. Nonfunctional features (shown in Fig. 7) are features provided by the middleware in terms of QoS, for example, security, fault tolerance, reliability, real-time support, etc.

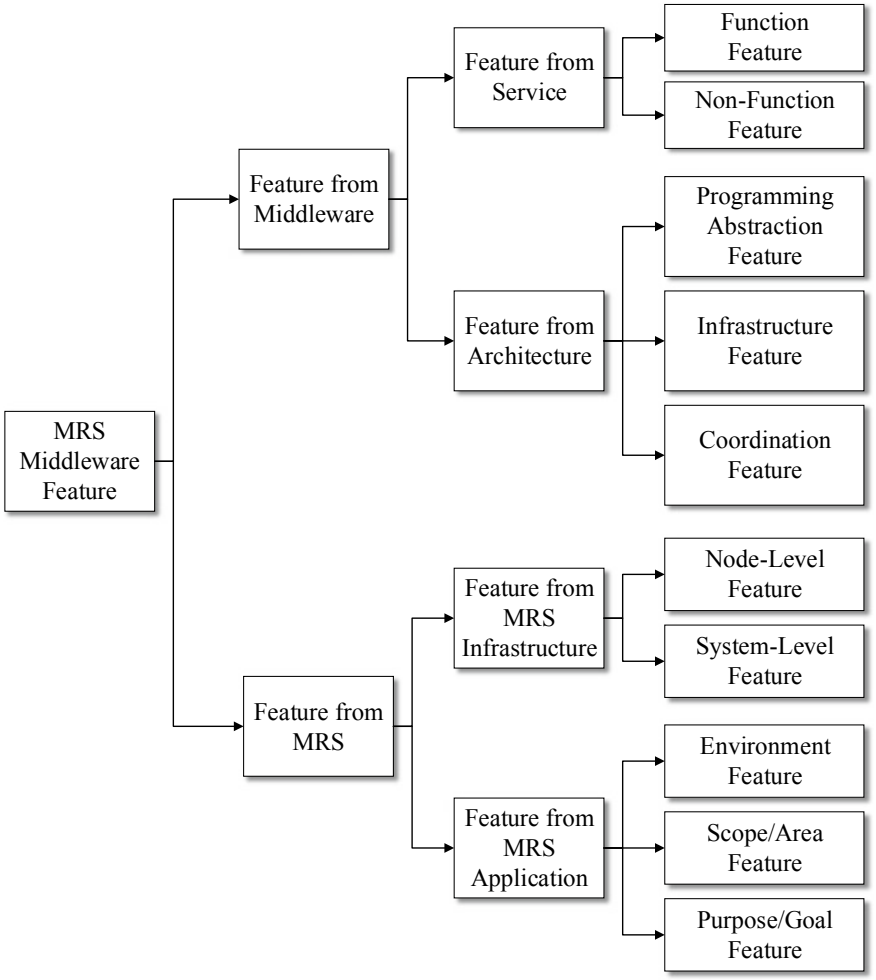


Fig. 5 Overview of feature tree of MRS middleware

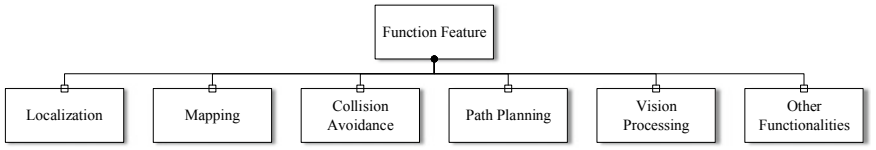


Fig. 6 Function feature

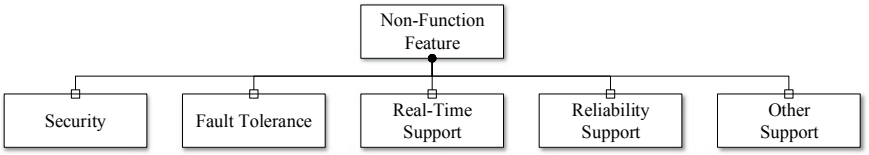


Fig. 7 Nonfunction feature

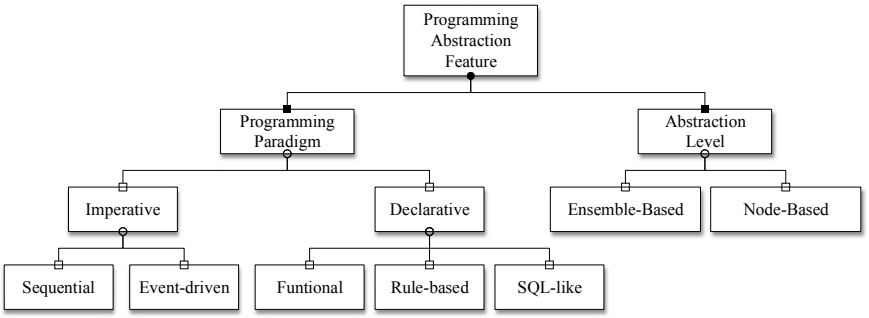


Fig. 8 Programming abstraction feature

Holonomic MRS middleware provides as many nonfunctional features as possible for developers so that they can choose a set of the features depending on specific applications. You cannot have your cake and eat it too, it is very suitable for some of the nonfunction features. For example, if the developer desires the privacy features, it may consume more time and consequently affects the real-time feature. Similar issue may be observed when fault tolerance and real-time support are provided by the middleware. There is always some form of trade-off between different nonfunction features. Such conflicts of nonfunction features are ubiquitous for middleware in other fields too such as wireless sensor network [19] and cloud computing [16].

Modern MRS middleware always provides a programming model or programming abstraction to facilitate development. A programming model masks the complexity of the system. Programming paradigm and abstraction level serve as the two fundamental elements of a programming model (shown in Fig. 8). The programming paradigm refers to the abstractions used to represent individual elements of a program. The individual elements of a program include constants, variables, clauses (iterations, conditions, etc.), and functions. Programming paradigm of a programming model can be imperative or declarative. While programming with imperative approach, the state of the program is explicitly expressed through statements. Relevant subcategories of imperative approaches include sequential and event-driven. On the other hand, while using a declarative programming model, the application goal is described without specifying how it is accomplished. Declarative approaches can be further classified into functional, rule-based, SQL-like, and special-purpose. The abstraction level refers to how developers view the multi-robot system and can be either node-based or ensemble-based. Node-based abstraction is used in traditional

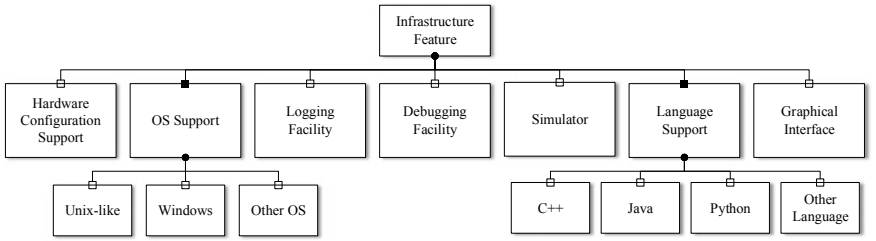


Fig. 9 Infrastructure feature

programming model where each robot is programmed, respectively. When a group of robots is assigned a task, it is natural to think about what the robot ensemble should do. This leads us to consider the ensemble-level abstraction. The entire MRS can be viewed as a single and monolithic unit for the programmer. Ensemble-level abstraction is referred to as macro-programming in wireless sensor networks [35].

There is a wide range of infrastructures (shown in Fig. 9) for MRS middleware. Infrastructures of MRS middleware include hardware configuration support, operating system support, logging facility, debugging facility, simulator, language support, and graphical interface.

- Since MRS middleware may be applied to all kinds of robotic system, hardware configuration support is required to configure the hardware of a specific kind of robot.
- MRS middleware must support a specific operating system or be cross-platform. Traditional operating systems can be UNIX-like OS, Microsoft Windows, Java virtual machine, and others.
- Logging facility and debugging facility are essential and useful for application development, algorithm evaluation. Looking up the log and debugging information, developers can have direction for development and improvement, which save significant amount of time.
- Simulator is useful when deployment is costly, hardware is unavailable, or developers want to testify algorithms before deployment.
- Language support is another necessary feature for MRS middleware. It can support one or several languages, for example, C++, Java, Python, etc.
- Graphical interface can be used to visualize the MRS and for human–robot interaction purpose.

Coordination (shown in Fig. 10) is a general issue in multi-agent system as well as in MRS where each realistic robot is regarded as an agent. A robot is a computational device capable of sensing, computing, and locomoting. The sequence of sensing, computing, and locomoting form a computation cycle of a robot. The coordination method is classified based on the relationship among computation cycles of the robots. In the asynchronized setting, the robots in the MRS do not have a common notion of the time. That is to say, there is no assumption on the relationship among the cycles of the same robot or different robots. The only assumption is that all

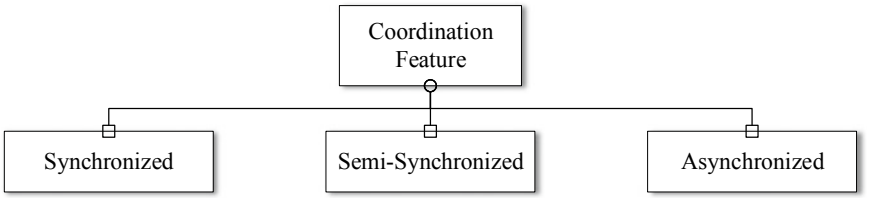


Fig. 10 Coordination feature

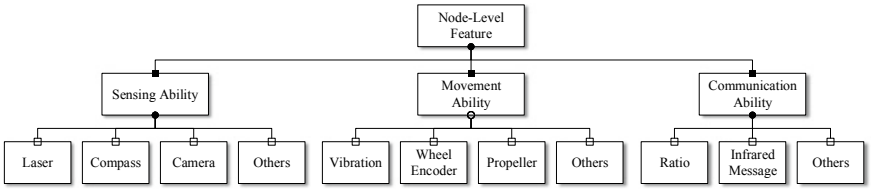


Fig. 11 Node-level feature

cycles finish in finite time. The robots are said to run in a semi-synchronized setting if all robots share a global clock and their actions are atomic. The robots can be either active or inactive at each clock tick and only robots in active state perform their cycles. To make sure every clock tick and every robot to be meaningful, it is restricted that at least one robot is active at every clock tick and every robot becomes active for infinite time instants. In a special case, every robot is active at every time instant. In this case, the robots are said to be fully synchronized. In this setting, all the robots are in the same state at each clock tick.

With respect to node-level features (shown in Fig. 11) of a single robot, it refers to the hardware features relating to sensing ability, locomotion ability, computation ability, and communication ability. For the sensing part, an individual robot may contain laser, compass, camera, microphone, etc. For the locomotion part, each robot may use vibration, wheel encoders, or propellers to navigate the environment. For the computation part, the CPU frequency, the memory size, and the data storage size vary a lot. For the communication part, ratio, infrared signals, Wi-Fi, and Bluetooth can be utilized to achieve it.

In terms of system-level features (shown in Fig. 12) of the whole MRS, it can be categorized by coordination method, embedded network protocol, and communication model. For the coordination method in an MRS, it can be centralized where there is a central controller, decentralized where the MRS is divided into groups, or fully distributed where all robots are equal. For communication, robot network may utilize TCP, UDP, ZigBee, or other network protocol. Communication is a general issue in network systems as well as in MRSs which can be regarded as robot networks. Communication features can be classified in the light of awareness, scope, and addressing. Awareness feature within communication can be further classified into explicit or implicit. If the communication is explicitly exposed to developers,

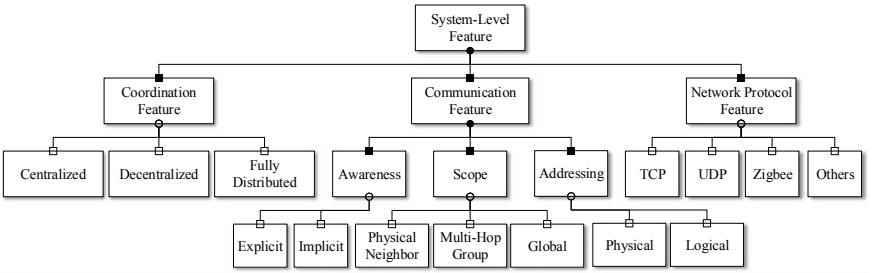


Fig. 12 System-level feature

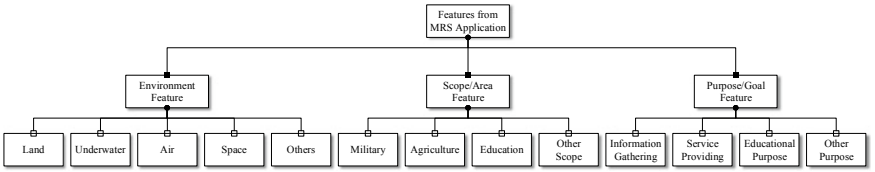


Fig. 13 Features from MRS application

it is termed as explicit. On the other hand, if the communication is hidden behind some higher level construct, it is said to be implicit. The scope of communication refers to the set of robots that exchange data to accomplish a given task. Physical neighborhood, multi-hop group, and system-wide serve as the three approaches for the scope of communication. The scope is physical neighborhood if programmers are only provided with method for exchanging data among robots within direct radio range. The scope is said to be multi-hop group if data exchange can be achieved with using multiple-hop transmission. The scope is system-wide if all the robots in the MRS are possible to be involved in data exchange. With respect to addressing in MRS middleware, it utilizes physical addressing if the target robots are identified using statically assigned identifiers. Otherwise, the target robots are identified through properties provided by the applications. This kind of addressing approach is logical addressing, which is generally called attribute-based addressing in wireless sensor networks [2].

In terms of application features, we sort it using environment, scope/area, and purpose/goal, respectively (shown in Fig. 13). An MRS can be deployed in one or more kinds of environments including land, underwater, air, and even space environment. The area of MRS application can be military, agriculture, education, household, manufacturing, etc. The purpose of an MRS application can be information gathering, service providing, educational usage, and others. Full discussion of the MRS applications can be found in Sect. 2.2.

5 Representative Middleware for MRS

In this section, we give a comprehensive overview of some popular middleware for MRS. This list is not exhaustive and there are many more middleware such as OROCOS [13], SmartSoft [84], CLARAty [93], etc. that have not been discussed in this work. We have focused on more recent and popular middleware for robotics. For each middleware that has been included below, we have discussed the architecture, objective, development tools and utilities provided, and platforms and programming languages supported by the middleware. An overview of various MRS middlewares, discussed in this section, has been shown in Table 2. This overview is done on the basis of middleware-specific features illustrated in Fig. 5, which shows the overview of feature tree of MRS middleware.

Player/Stage: Player[5] is device server that provides clients with programming interfaces to control robots comprising of sensors and actuators [31]. Player is implemented in C++ as a multi-threaded TCP socket server for transparent robot control. Socket-based robot server provides many benefits such as platform independence, language independence, and location neutrality which means a client can access and control robotic devices anywhere on the network. Player has been designed to support heterogeneous devices and clients simultaneously at different timescales [31]. One-to-many client/server architecture has been followed which implies that one server can serve multiple clients. Each client is connected to Player by a TCP socket connection, while a device can be connected to Player by any appropriate method. Client can be implemented in any language providing socket mechanism such as C, C++, Tcl, Java, Python, etc.

Player is modular; therefore, devices can be added dynamically. UNIX model of treating devices as files has been chosen to provide an abstraction for a variety of devices. To receive sensor readings, client opens the device with read access while for controlling an actuator, client must open the device with write access. Each device has an associated command and data buffer that provides an asynchronous communication channel between device threads and client reader and writer threads. Clients and devices are decoupled from each other. Player also supports request–reply mechanism, similar to `ioctl()`, for configuration requests that can be used to access specific hardware features. There is no device locking mechanism implemented in Player; therefore, clients can overwrite commands of the other clients.

Stage is a simulator that is used for simulating population of mobile robots, sensors, and environmental objects. This enables development and testing of clients without accessing real hardware and environment. Stage simulator is also useful for experimentation of novel devices that have not been developed yet [31]. Sensors and actuator models in Stage are available through normal Player interface. Usually, clients cannot differentiate between real and simulated stage equivalents. Stage also supports non-locking, platform independence, and language independence characteristic of interfaces in Player.

[5] http://playerstage.sourceforge.net/.

Table 2 Overview of existing MRS middleware

Middleware	Function features	Nonfunction features	Programming abstraction	Infrastructure features	Coordination feature
Player	Localization, mapping, path planning, collision avoidance, vision processing	Security, robustness	Imperative, node-based	Unix-like/Windows, logging/debugging facilities, Simulator—Stage/Gazebo, C, C++, Java, Python, etc.	Asynchronized
Orca	Localization, mapping, etc. reusable from other projects	None	Imperative, node-based	Unix-like/Windows/Mac OS X, logging/debugging facilities, graphical interface: GOrca, C++, Java, Python, PHP, C#, and Visual Basic	Asynchronized
Miro	Mapping, localization, path planning, collision avoidance, speech recognition, vision processing	Robustness	Imperative, node-based	Unix-like/Windows, logging/debugging facilities, graphical interface: Qt GUI, C++	Asynchronized (Event-driven control)
MIRA	Localization, mapping, path planning, collision avoidance, vision processing	Security, robustness, reliability, fault tolerance, real-time support	Imperative, node-based	Linux/Windows, logging/debugging facilities, graphical interface: Qt GUI, C++, Python, JavaScript, etc.	Asynchronized
OpenRDK	Mapping, localization, path planning, navigation, collision avoidance, vision processing	Robustness	Imperative, node-based	Unix-like, logging/debugging facilities, simulator: USAR-Sim/Stage/Gazebo, C++	Asynchronized

Table 2 (continued)

Middleware	Function features	Nonfunction features	Programming abstraction	Infrastructure features	Coordination feature
MARIE	Mapping, localization, path planning, collision avoidance, vision processing	Robustness	Imperative, node-based	Unix-like, logging/debugging facilities, simulator: Stage/Gazebo, graphical interface: logviewer, C++	Asynchronized
Urbi	Vision processing	None	Imperative, node-based	Unix-like/Windows/Mac OS X, logging/debugging facilities, simulator: Webots, graphical interface: UrbiLab, C++, Java, MATLAB, Python, etc.	Both sychronized and asychronized
MRDS	Speech recognition, vision processing	Security, fault tolerance, robustness	Declarative, ensemble-based	Windows, logging/debugging facilities, simulator: Visual Simulation Environment, graphical interface: Visual Programming Language, C#, Visual Basic, and Iron Python	Asynchronized

Table 2 (continued)

Middleware	Function features	Nonfunction features	Programming abstraction	Infrastructure features	Coordination feature
RoboComp	Mapping, localization, etc. reusable from other projects such as Player, Orca, and ROS	None	Imperative, node-based	Unix-like/Windows/Mac OS X, logging/debugging facilities, simulator: Stage/Gazebo, graphical interfaces: managerComp, monitorComp, etc., C++, Java, Python, Ruby, C#, PHP, and Objective C	Asychronized
ROS	Mapping, localization, etc. reused from other projects such as Player, OpenRAVE, etc.	Real-time support, robustness	Imperative, node-based	Unix-like/Windows (partial)/Mac OS X, logging/debugging facilities, simulator: Stage/Gazebo, graphical interfaces: rxplot, rxgraph, C++, Python, Octave, and LISP	Asynchronized
WURDE	Localization, collision avoidance, vision processing	Robustness	Declarative, node-based	Unix-like, logging/debugging facilities, simulator: Stage, graphical interface: RIDE, C++	Asynchronized

Table 2 (continued)

Middleware	Function features	Nonfunction features	Programming abstraction	Infrastructure features	Coordination feature
OPRoS	Mapping, localization, path planning, collision avoidance, vision processing	Fault tolerance, robustness, real-time support (under development)	Declarative, node-based	Linux/Windows, logging/debugging facilities, simulator: OPRoS Simulator, Robot Builder, graphical interface: Component Composer, C++, Java	Asynchronized (event-driven)
RT-Middleware	Mapping, localization, path planning, collision avoidance, vision processing	Real-time support, robustness	Imperative, node-based	Unix-like/Windows, logging/debugging facilities, simulator: Stage, graphical interface: RTSystemEditor, C++, Java, and Python	Both synchronized, and asynchronized
ASEBA	Mapping, localization, path planning, collision avoidance, vision processing	Real-time support, robustness	Imperative, node-based	Unix-like/Windows, logging/debugging facilities, simulator: Enki Simulator, graphical interface: ASEBA IDE based on Qt4, ASEL (ASEBA Event Scripting Language)	Asynchronized (Event-based)

Since Player is freely available as open-source, many improvements have been done since the original version [32]. Some major improvements related to simplicity and flexibility have been done in [22]. Player has been divided into two parts, the core and transport layer. Separation of Player core from transport layer provides more flexibility. Original version [32] was a TCP-based device server, but now it can support many other configurations. Other transport layers, or no transport layer, can also be used. Player is supported on most of the UNIX flavors and on Windows using Cygwin.

Orca: Orca[6] is a framework that can be used for the development of component-based robotic systems. Complex robotic systems can be developed by piecing together the components provided by Orca. The main objective of Orca is to promote software reuse. Orca does not impose any constraint on the component granularity (size of modules used to make up the complete system), system architecture (any architecture such as centralized, blackboard, strictly-layered, strictly-decentralized, or mixed can be implemented), interfaces, and component architecture [56].

Orca uses Internet Communication Engine (Ice) for communication between interfaces [56]. Slice, a specification language for Ice, is used for defining interfaces. There are many Ice services such as IceGrid Registry, IceGrid Node, IceBox, and IceStorm which are extensively used in Orca. IceGrid Registry is a centralized registry for naming service. IceGrid Node is a software activation service. IceBox is an application server that is responsible for starting and stopping of application components. Application components are deployed as a dynamic library which makes them easy to deploy and administer, and also optimizes the communication between components within the same application server. IceStorm is an event service which forward the messages received from a server to multiple clients without marshaling or demarshalling them. IceStorm can also weaken client dependencies by configuring multiple threads.

Orca also provides a library called libOrcaIce which provided simplified API that can be used for development of robotic applications [56]. This lowers the barrier for developers as the majority of functionalities used for robotic applications are provided by Orca library. To allow the use of Orca on wider platforms, CMake is used to build system. Orca can be used on different operating systems including Linux, several flavors of Windows, and Mac OS X. Programming languages that are supported are C++, Java, Python, PHP, C#, and Visual Basic. Also, Ice Client and server are language independent so they can be implemented in any programming language.

Miro: Miro[7] is a three-layered middleware for mobile robot applications which is designed and implemented using object-oriented approach [92]. The three layers from bottom to top are MIRO device layer, MIRO service layer, and MIRO class framework layer. The higher layers access lower layers using interfaces. MIRO device layer is a platform dependent layer that provides classes to interface and abstracts the low-level sensors and actuators within a robot. The classes also allow access to

[6] http://orca-robotics.sourceforge.net/index.html.

[7] https://sourceforge.net/projects/miro-middleware.berlios/.

low-level hardware resources using ordinary method calls. MIRO service layer provides service abstraction for sensors and actuators with event-based communication by using CORBA interface definition layer (IDL). The services are implemented as network transparent CORBA objects which enable language and platform independence. Sensors and actuators are presented in a platform-independent manner by the use of classes in this layer. MIRO class framework layer provides functional modules such as mapping, localization, behavior generation, logging and visualization facilities, etc. which are extensively used for mobile robotic application development. Besides providing common functionalities for application development, MIRO class framework also provides functionality for experimental evaluation.

All MIRO functionalities have been implemented in C++. The communication mechanism is developed using TAO package which is an implementation of CORBA-based on adaptive communication engine (ACE). Client–server model has been used for communication between objects. MIRO implementation includes three types of clients (sample client, test client, and monitoring client) for testing and evaluation of service functionalities. Apart from event-driven communication between services, synchronous and asynchronous communications are also used.

MIRA: MIRA[8] is a decentralized middleware that supports the development of fully distributed robotic applications. The objective of MIRA is to support the development of real-world applications; therefore, mechanisms have been used to address issues such as memory consumption, latency, fault tolerance, and robustness. Each application is composed of different processes that can be located on different machines. Each process further consists of multiple software modules called units which implement algorithms to solve any task. In case multiple units are present in a single process, then each of the units runs in its own thread.

MIRA is written in C++ but it can be interoperable with other programming languages such as Java, Python, etc. Reflection and serialization are two concepts that have been widely used in implementing different mechanisms in MIRA. MIRA uses a reflect method to reflect and serialize any arbitrary class since reflection and serialization are not supported natively by C++. This mechanism enables the complete use of object-oriented programming paradigm. This allows transport of not only simple data but also complex objects including robot models, GUI components, etc. to the remote side. Use of serialization makes MIRA interoperable with other programming languages and middleware. Serialization formats such as XML, JSON, and binary are currently supported by MIRA.

Two communication mechanisms are supported by MIRA, message passing, and remote procedure calls (RPC). There is no central server used in MIRA for name look or other management tasks. MIRA supports robustness and reliability by using peer-to-peer architecture for communication between different processes. Communication between units and message exchange is done using named channels. Channels allow one-to-one, one-to-many, and many-to-many communication [27]. MIRA supports autonomous handling of multi-threading and data synchronization. Slot-based communication avoids unnecessary copying and blocking of data when there

[8] http://www.mira-project.org/joomla-mira/.

is simultaneous read and write access. Slot-based communication also helps in reducing memory usage. MIRA reduces latency of RPC by using futures, which act as proxy for the result of asynchronous calls. MIRA is currently supported for Linux and Windows operating system.

OpenRDK: OpenRDK[9] is a modular software framework designed to develop distributed and mobile robotic applications. The objective of OpenRDK is to support modularity and code reusability of software to enable easier and faster development of robotic application. The main entity of the software framework is a software process called agent. Single thread inside an agent process is called module, which can be loaded and started dynamically once agent is running [17]. An agent configuration is a list of modules that are to be loaded and executed, their interconnection layout, and value of their parameters. All modules publish the data they want to share in a repository. Variables published by modules, i.e., input, output, and parameter, are called properties. Each property is assigned to a globally unique URL address. These URL addresses enable modules to transparently access modules within the same agent or remotely. There are some special queue objects that are also addressed using global URL just like other local properties.

OpenRDK uses multiple processes with multiple threads. Since information sharing between modules is done with the help of URL, it introduces some coupling between modules which adversely affects the modularity of the whole system. To resolve this issue, property links are specified in configuration file that allows modules to refer to different names for the properties, thus avoiding any coupling between modules. While sharing of information within the same agent can be done using repository, inter-agent information sharing can be done by either property sharing or message sending. OpenRDK also contains RConsole which is a graphical tool used for remote inspection and management of modules. Other modules for connecting to simulators or for logging are also provided by OpenRDK. OpenRDK is written in C++. It can run on a UNIX-like operating system. OpenRDK does not focus on platform independence of the software framework.

MARIE: Mobile and Autonomous Robotics Integration Environment (MARIE)[10] is a distributed component-based middleware framework designed to develop robotic applications by enabling integration of new and existing systems [23]. The objective of MARIE is to enable software reuse, support multiple sets of concepts, and support a wide range of communication protocols, mechanism, and robotic standards. MARIE supports multiple levels of abstraction by utilizing layered software architecture consisting of three layers, which are core layer, component layer, and application layer. Core layer is the lower level layer that provides tools for low-level functionalities such as communication, distributed computation, data handling, and many low-level operating system functions. Component layer implements a framework to add components and support domain-specific concepts. Application layer consists of tools required to build robotic applications from available components.

[9] http://openrdk.sourceforge.net/index.php?n=Main.HomePage.

[10] http://marie.sourceforge.net/wiki/index.php/Main_Page.

MARIE uses mediator interoperability layer (MIL) to act as a mediator for interaction with each component independently. MIL is implemented as virtual space where components can interact with each other using a common language. This design leads to the decoupling between components, increases reusability, interoperability, and reduces the complexity of managing a large number of centralized components. MIL is composed of four types of components, which are application adapter (AA), communication adapter (CA), application manager (AM), and communication manager (CM). AA is responsible for interfacing applications with MIL. CA is responsible for communication between components by adapting different communication mechanisms. AM is responsible for management of all components in the system, and CM is responsible for management of communication between AAs.

MARIE has been written in C++. MARIE does not focus on any specific communication mechanism; rather it uses communication abstraction framework, called port, for provided communication protocols and component interconnection. It uses adaptive communication environment (ACE) library to implement for transport layer and low-level operating system function implementation.

Urbi: Urbi[11] is a software platform for developing robotic applications. Urbi is based on an event-driven scripting language, URBISCRIPT, and distributed component architecture [8]. URBISCRIPT is designed not only to create robotic applications and controlling robots but also it is the foundation of the communication protocol based on which client/server architecture for Urbi software platform is built. Multiple clients can interact concurrently with a server by means of URBISCRIPT. Urbi server which lies above the operating system is responsible for abstracting low-level hardware details. Urbi platform interacts with underlying operating system using engines which are also responsible for running the Urbi server. Urbi kernel is another part of the Urbi server that provides primitive services including urbi virtual machine (UVM). URBISCRIPT running on top of urbi virtual machine (UVM) is responsible for providing CPU independence.

Diversity in robots is addressed by the use of UObject architecture. UObject enables communication between low-level and high-level components, and their interaction with URBISCRIPT. Complex data flowing between multiple components can also be handled by the use of UObject API. UObject can be either plugged into the server or also used as standalone remote process. Besides low-level abstraction provided by UObject, high-level abstraction is also provided by using Urbi naming standard. These abstractions enable development of portable applications.

There are many graphical applications such as UrbiLive and UrbiLab provided by Urbi platform to enable easy interaction with robots. UrbiLive is a graphical editor useful for composing and chaining actions based on external events. UrbiLab is a graphical tool used as Urbi server inspector and effector. UrbiLab can also be used for remote control of robots. Urbi is open to programming environments such as Java, MATLAB, and Python. Although Urbi platform is based on C++ and URBISCRIPT, it is not necessary to know these languages to program robots and components.

[11] http://www.gostai.com/products/urbi/.

Microsoft Robotics Developer Studio (MRDS): Microsoft Robotics Developer Studio (MRDS)[12] is a service-driven robotic studio that follows representational state transfer (REST) pattern [40]. Decoupled software services are used for interaction with robots. Use of decoupled services enables modularity and code reuse. Services are used for both robot interaction and implementation of functionalities such as web-based error reporting, wireless communication, etc. The interaction between services is done by the use of XML-based configuration manifest file. Manifest file enables start-up of services by MRDS by defining partnership between services. The partnership of services also enables registration between services, message passing, and fault notification. Partnership and distributed messaging are enabled by the use of software library called Decentralized Software Services Protocol (DSSP). MRDS also uses another software called Coordination and Concurrency Runtime (CCR) for handling state updates and message processing. CCR also enables abstraction of complex functionalities such as memory locking and communication between various operating systems. There are two other main components in MRDS, which are visual programming language (VPL) for graphical interface and visual simulation environment (VSE) for running simulation.

There are some utility services provided by MRDS. A control panel enables the user to view all currently running services and links between them. There is a message logging service that runs built-in filtering to provide a debugging view of the system. Resource diagnostic service is also provided to enable the developer to obtain additional debugging and performance evaluation information. A 3D simulator, based on Microsoft DirectX technology, is also included in MRDS. The simulator is used for both graphics and physics simulations. MRDS is implemented in .NET. Services in MRDS can be written in any .NET compatible language. Service implementations have been done in C#, Visual Basic, and Iron Python. Simple object access protocol (SOAP) interface can also be used to interface services with other programming interfaces. MRDS is a popular proprietary middleware that only supports Windows platform.

RoboComp: RoboComp[13] is a component-oriented robotic framework that focuses on ease of use and rapid development of robotic applications [57]. Robocomp is based on Ice which is extended further by the use of different classes and tools. Components used in Robocomp consist of three main elements, which are server interface, worker class, and proxies that are used for communicating with other components. Worker class implements the core functionality of components. Server interface and worker class run in different threads to avoid delays. There is another optional common interface called CommonBehavior that is used for accessing the parameters and status of components.

Different tools provided by RoboComp are used for providing functionalities such as monitoring, management, debugging, simulation, etc. These tools are as follows:

[12] https://www.microsoft.com/en-us/download/details.aspx?id=29081.

[13] http://robocomp.github.io/website/.

(a) componentGenerator: This tool makes the task of the programmer easier by automatically generating the skeleton of the new component and even the code pieces for the programmer.
(b) managerComp: managerComp is a graphical tool that can be used for building and running the system. Both local and remote components are managed can be managed by use managerComp. This tool also makes use of CommonBehavior interface to visually access the parameters of the components.
(c) monitorComp: This tool is used for connection and monitoring of components. monitorComp provides a graphical interface for testing the components in an easier way. Testing is done either by the use of custom monitoring code or template available to test HAL components.
(d) replayComp: This tool records the output of components to replay them. This is also a graphical tool that is useful for debugging purposes.
(e) Simulation Support: RoboComp makes use of two widely used open-source simulators, Stage and Gazebo, for simulation purpose.
(f) loggerComp: This tool is used for analyzing the execution and interaction of components. This tool also provides a graphical interface to display different types of information.

RoboComp can be deployed on any computer system supporting Ice. Platforms supported by RoboComp are Linux, Windows, Mac OS X, Android, and iPhone. Any programming language that supports Ice can be used for RoboComp, which includes C++, Java, Python, C#, Ruby, PHP, and Objective-C.

ROS: ROS[14] is a modular framework for developing robotic systems [74]. ROS provides a structured communication layer above operating system of the host. Multiple processes running in a system are connected using peer-to-peer topology instead of using a central server. ROS is language-neutral. A simple and language-neutral interface definition language (IDL) is used to provide support for cross-language development. All the functionalities in ROS are developed using small modules. Use of modular architecture reduces complexity and enhances stability. All the driver and algorithm development is done in standalone libraries which are independent of ROS. Small executables are created inside the source code which exposes library functionality to ROS. This mechanism makes code reuse easier and helps in unit testing.

There are some fundamental concepts that have been defined for ROS implementation, which are nodes, messages, topics, and services. A system is composed of multiple nodes which are processes that perform computation [74]. Communication between nodes is using messages. A message can consist of other messages, or array or messages. Topic-based publish-subscribe communication paradigm has been used. To enable request/response communication, the concept of services has been defined. There can be multiple simultaneous publishers and/or subscribers for a single topic, and a single node can publish and/or subscribe to multiple topics.

ROS follows tool-based design, and thus there are many tools provided with ROS for different scenarios. ROS uses rconsole library to enable logging and monitoring

[14] http://www.ros.org/.

of distributed system. Packaging functionality is enabled by the use of roslaunch tool. Collaborative development is enabled by the use of utilities such as rospack and rosbash. ROS uses a utility named rostopic for filtering messages. There are tools such as rxplot and rsgraph that are used for generating plots and graphs. ROS has been designed to be language-neutral. ROS supports four different types of programming languages, which are C++, Python, Octave, and LISP.

WURDE: WURDE (Washington University Robotics Development Environment) is a modular middleware for developing robotic systems [37]. The objective of WURDE is to develop middleware that is easy to use even for beginners. WURDE provides set of abstraction and utilities to achieve this objective. Four layers of abstraction are provided by WURDE, which are communication, interface, application, and system. WURDE does not use specific communication protocol instead the communication layer defines types and methods for moving data to communication adapter [28]. Interface layer describes the data required by each type of robot. Interfaces are described using XML. Application layer provides API for controlling different aspects of applications. System layer is topmost abstraction layer which is used for connecting different applications.

WURDE does not use any specific software architecture; instead, the robotic system is developed as system of small interconnected modules. WURDE uses asynchronous communication for communication between modules. One of the modules provided by WURDE is a tasking and control interface called Robot Interaction Display Environment (RIDE). RIDE enables single user to control and task multiple robots at same time. Implementation of WURDE is not yet complete as there are many software packages that are still needed to be developed.

OPRoS: Open platform for robotic services (OPRoS)[15] is a distributed component-based platform for developing robotic systems [41]. The objective of OPRoS is to enable full development of robot software by providing required developmental tools, middleware services, component execution engine, simulation environment, and other utilities. Robotic service in OPRoS is composed of loosely coupled components. There are two types of components, atomic component and composite component. Components can have different granularities. Communication between components is done by the use of ports on each specific component. Each component can have multiple ports that are used for transmission of different types of information such as method invocation, data, and events. Components in OPRoS can support either of the three different types of execution modes, which are periodic, nonperiodic, and passive.

All the information related to components such as port types, execution semantics, properties, and other relevant information is stored in an XML file called component profile [41]. There are other XML profiles such as service profile, data profile, and application profile. Component and application profile is used by communication execution engine for execution and management of components. Component execution engine provides abstractions to developers by hiding low-level details such as thread management, resource allocation, and other functions offered by the oper-

[15] http://www.opros.or.kr/display/opros/OPRoS+Wiki.

ating system. Component execution engine also provides a self-configurable fault tolerance module for detecting faults and anomalies.

OPRoS also provides development tools for authoring atomic components and composing components. Component authoring tool, as the name suggests, is used for authoring atomic components. This tool also supports debugging, execution control, and monitoring of atomic components. Component composer is used for composing components to develop the robotic application. Component composer can also be used for remote control and monitoring of multiple component execution engines concurrently. All these tools can be supported on any operating system where eclipse is installed. OPRoS currently supports both Windows and Linux operating systems.

RT-Middleware: RT (Robot Technology)-Middleware[16] serves as a distributed component-based middleware for developing robotic systems [4]. It studies modularization of robotic elements and proposed RT-Components as the basic software unit based on Common Object Request Broker Architecture (CORBA). RT-middleware supports the construction of various networked robotic systems by the integration of various RT-Components. An open-source implementation called OpenRTM-aist [5] was developed for feedback from the robotic research community.

An RT-Component is composed of a component object as the main body, activity as the main process unit, and input ports (InPorts) and output ports (OutPorts) as data stream ports. The activity serves as a controller of the device and processes the designed tasks continuously. The activity has eleven states, and each state is possible to have three methods called entry, do, and exit. These three methods will be called automatically on entry to, being in, and on exit from the state, respectively. The transition of the states is uniform for all RT-Components. Hence, developers may implement the methods for each state to build a new RT-Component.

The InPorts and OutPorts take advantages of publisher/subscriber model. On the one hand, an InPort serves as a subscriber and may subscribe several OutPorts. It also provides a common method called `InPort::put` to allow data to be written. On the other hand, an OutPort serves as a publisher and write data to those InPorts who have subscribed it by using the method of `InPort::put`. It also provides several subscription types, e.g., Once, Periodic, and Triggered.

The RT-Middleware provides several methods for integrating RT-Components. These methods include assembly GUI tool, script language, XML file, other RT-Components, and other application programs. By integration of RT-Components, applications can be built from bottom to up. Such applications include network distributed monitoring system [42] and intelligent home service robotic system [43].

ASEBA: ASEBA[17] is an event-based middleware supporting distributed control and efficient resources exploitation among multiple microcontrollers in a robot [52, 55]. The ASEBA is specially designed for robots with more than one microcontroller sharing a bus for communication.

The ASEBA abandons traditional architecture where the main microcontroller controls all the other microcontrollers and manages all data transfers. As an event-

[16] http://openrtm.org.

[17] http://mobots.epfl.ch/aseba.php.

based middleware, it utilizes multi-master bus in which all microcontrollers can initiate data transfers. The multi-master bus enables asynchronous messages, called events, transferred between microcontrollers. Without control of a centralized main processor, load on bus is significantly reduced. Also, the main process can get released from processing messages and be dedicated for CPU-intensive tasks such as path planning and image processing.

A scripting language called AESL (ASEBA Event Scripting Language) is provided to describe even emission and reception policy in ASEBA. A piece of AESL program can be compiled into bytecode using the designed compiler and ran on the implemented virtual machine. The compiler, together with a script editor and a distributed debugger, forms into the integrated development environment for ASEBA. The virtual machine is lightweight with less than 1000 lines of C program including debugging logic.

The ASEBA has been successfully deployed in the handbot for the task of climbing a shelf and in the marXbot to improve the performance of behaviors in terms of a polling-based approach. The ASEBA has been utilized for the purpose of education [53] and managing a collection of single-microcontroller robots [54].

6 Future Directions and Challenges

In this section, we have given some observations regarding future directions and challenges for MRS middleware. These observations have been made after reviewing existing middleware for MRS. We discuss which design goals have been commonly addressed by existing middleware and which ones need more research effort. Since design of middleware is dependent on hardware, it is important to develop multi-robot systems, which are cheaper, smaller, and have better sensing, processing, and communication technologies. Due to the limitation in technology, each robot in MRS has very limited processing power, storage, communication capability, and battery capacity. These resource limitations make it difficult to develop sophisticated middleware for MRS. One of the major research directions is to develop lightweight middleware that can be used to enable different design goals. Another challenging issue arises due to the heterogeneity of robots. MRS work by collaborating and coordinating with each other, however, heterogeneity of robots and communication protocols makes it a challenging task. Collaboration involves partitioning and distribution of complicated tasks among multiple robots; therefore, middleware should provide lightweight algorithms for task partitioning and management. Scalability will be another major challenge in coming future. Middleware should be designed to support scalability and dynamic configuration of system.

Middleware evaluation is a major part in analyzing the middleware. Middleware is usually evaluated by quantifying the system parameters based on some application example; however, there are some issues with this approach. Application examples that are used for middleware evaluation only focus on some specific system requirements which means evaluation depends highly on the application example being

used [69]. Besides, different middleware architectures are developed for different applications, and thus it may not be the best criteria to evaluate a middleware based on some specific parameters. Another issue with middleware evaluation is that it is difficult to quantify the usability of middleware [69]. Several mechanisms such as the use of a questionnaire, number of lines of code, or time to develop the system have been used to determine whether middleware is easy to use or not but, this is not very efficient. Middleware evaluation is currently a challenging issue for researchers and developers which requires more research efforts.

Robot software architecture can usually be classified into three types, which are object-oriented robotics, component-based robotics, and service-driven robotics [1]. Most middleware examples that have been discussed in Sect. 5 use small modules or components for developing a robotic system. Very few middleware, in general, follow monolithic approach for developing robotic software platform. It is preferred to use small modules for system development as it reduces the complexity of the whole system and enables reusability. Component-based approach is a common choice for building middleware for MRS. Orca, MARIE, Urbi, Robocomp, OPRoS, and RT-Middleware follow component-based architecture for developing middleware. Miro and Mira follow object-oriented paradigm, and MRDS is a service-driven middleware that follows REST pattern. All the other remaining middlewares such as ROS, WURDE, OpenRDK, Player, etc. use modular architecture. Out of all these middleware architectures, ROS and Player are most popular among robotic system developers. Although component- and modular-based approaches offer many benefits, a challenging issue is to integrate different components which results in issues related to communication, interoperability, and configuration [65]. Currently, a new trend is emerging in robotics community to use model-driven engineering for robotics software [1]. SmartMDSD Toolchain [89] is a toolchain based on model-driven software development approach that provides an integrated modeling environment to create an overall workflow for robotics software development.

We described some design goals for middleware in Sect. 3; however, we observed that each middleware is usually designed focusing on some specific goals. It is very difficult to achieve all design goals simultaneously [86]. Future work in MRS middleware should consider satisfying multiple design goals to enable multidimensional benefits. Out of all the design goals, flexibility and software reusability is the most common objective for existing middleware. Software reusability is enabled by the use of modular or component-based middleware architectures. Second observation is that most of the middleware are freely available as open source. This is usually done to enable debugging of the software and make it more popular. Since heterogeneity is a major issue in robotic system development, most middleware tries to provide some programming abstractions and make their system platform and language independent. Linux is the first choice of operating system that is supported by most middleware, Windows being the second, and there are some middlewares such as Orca and RoboComp that support Mac OS too. Most middlewares provide some graphical interface that enables management and monitoring of the robotic system. Orca, OpenRDK, MARIE, Urbi, MRDS, RoboComp, ROS, and OPRoS are some middleware examples that provide specific management and monitoring tools. Apart

from graphical interfaces, other mechanisms can also be used for management and monitoring purposes.

Although the features currently supported by existing middleware help in decreasing the complexity and provide some other useful features, there are many design goals which have not been fully addressed. One such design goal is collaboration among multiple robots which is currently achieved by existing middleware as they are usually modular and support distributed control. However, existing middleware does not focus much on providing some specific abstractions or tools to facilitate collaboration. A similar case is observed for dynamic resource discovery and configuration where specific tools and abstractions are not provided to enable dynamic configuration of system. Most middlewares do not provide explicit facilities to make the system more robust. Very few middlewares provide explicit fault tolerance capabilities. Out of all the middleware discussed, OPRoS, MIRA, and MRDS explicitly consider fault tolerance. Real-time support is another design goal which has not been addressed by many middleware. RT-Middleware, MIRA, ROS, and ASEBA are among the few middlewares that provide some form of real-time support. RT-Middleware provides some support for real-time processing, MIRA provides a mechanism to minimize latency, ROS enables real-time inspection and monitoring of any variable, and ASEBA also provides real-time support. Real-time support is essential for many robotic applications and thus, more research efforts are required to address this issue. Most middlewares currently do not focus much on supporting Quality of Service (QoS) requirements such as security, reliability, etc. Security and privacy are important concerns as MRSs are used for critical applications such as battlefield surveillance, exploration missions, etc. Very few middlewares provide security mechanism. Since MRSs are distributed in nature, it is challenging to develop distributed security algorithm. Although existing middlewares are developed to integrate sensors and actuators within the robots, integration with other technologies such as cloud computing and IoT has not been taken into much consideration.

Acknowledgements This work was supported by the ANR/RGC Joint Research Scheme [grant number A-PolyU505/12], the NSFC Key Grant [grant number 61332004], and the NSFC/RGC Joint Research Scheme [grant number N-PolyU519/12].

References

1. Ahmad, A., Babar, M.A.: Software architectures for robotic systems: a systematic mapping study. J. Syst. Softw. **122**, 16–39 (2016)
2. Akyildiz, I.F., Su, W., Sankarasubramaniam, Y., Cayirci, E.: Wireless sensor networks: a survey. Comput. Netw. **38**(4), 393–422 (2002)
3. Alimisis, D.: Educational robotics: open questions and new challenges. Themes Sci. Technol. Educ. **6**(1), 63–71 (2013)
4. Ando, N., Suehiro, T., Kitagaki, K., Kotoku, T., Yoon, W.K.: Rt-middleware: distributed component middleware for rt (robot technology). In: 2005 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 3933–3938. IEEE (2005)

5. Ando, N., Suehiro, T., Kotoku, T.: A software platform for component based rt-system development: Openrtm-aist. In: International Conference on Simulation, Modeling, and Programming for Autonomous Robots, pp. 87–98. Springer (2008)
6. Arai, T., Pagello, E., Parker, L.E.: Editorial: advances in multi-robot systems. IEEE Trans. Rob. Autom. **18**(5), 655–661 (2002)
7. Arbuckle, D., Requicha, A.A.: Self-assembly and self-repair of arbitrary shapes by a swarm of reactive robots: algorithms and simulations. Auton. Robots **28**(2), 197–211 (2010)
8. Baillie, J.C., Demaille, A., Hocquet, Q., Nottale, M., Tardieu, S.: The URBI universal platform for robotics. In: First International Workshop on Standards and Common Platform for Robotics (2008)
9. Beasley, R.A.: Medical robots: current systems and research directions. J. Robot. 2012 (2012)
10. Benitti, F.B.V.: Exploring the educational potential of robotics in schools: a systematic review. Comput. Educ. **58**(3), 978–988 (2012)
11. Broadbent, E., Stafford, R., MacDonald, B.: Acceptance of healthcare robots for the older population: review and future directions. Int. J. Soc. Robot. **1**(4), 319–330 (2009)
12. Bruce, J., Zickler, S., Licitra, M., Veloso, M.: Cmdragons: dynamic passing and strategy on a champion robot soccer team. In: IEEE International Conference on Robotics and Automation, 2008. ICRA 2008, pp. 4074–4079. IEEE (2008)
13. Bruyninckx, H.: Open robot control software: the orocos project. In: IEEE International Conference on Robotics and Automation, 2001. Proceedings 2001 ICRA, vol. 3, pp. 2523–2528. IEEE (2001)
14. Burgard, W., Moors, M., Fox, D., Simmons, R., Thrun, S.: Collaborative multi-robot exploration. In: IEEE International Conference on Robotics and Automation, 2000. Proceedings. ICRA'00, vol. 1, pp. 476–481. IEEE (2000)
15. Burgner-Kahrs, J., Rucker, D.C., Choset, H.: Continuum robots for medical applications: a survey. IEEE Trans. Robot. **31**(6), 1261–1280 (2015)
16. Calheiros, R.N., Ranjan, R., Beloglazov, A., De Rose, C.A., Buyya, R.: Cloudsim: a toolkit for modeling and simulation of cloud computing environments and evaluation of resource provisioning algorithms. Softw. Pract. Exp. **41**(1), 23–50 (2011)
17. Calisi, D., Censi, A., Iocchi, L., Nardi, D.: Openrdk: a modular framework for robotic software development. In: 2008 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 1872–1877. IEEE (2008)
18. Chalup, S.K., Murch, C.L., Quinlan, M.J.: Machine learning with AIBO robots in the four-legged league of robocup. IEEE Trans. Syst. Man Cybern. Part C (Appl. Rev.) **37**(3), 297–310 (2007)
19. Chen, D., Varshney, P.K.: Qos support in wireless sensor networks: a survey. In: International Conference on Wireless Networks, vol. 233, pp. 1–7 (2004)
20. Chitic, S.G., Ponge, J., Simonin, O.: Are middlewares ready for multi-robots systems? In: International Conference on Simulation, Modeling, and Programming for Autonomous Robots, pp. 279–290. Springer (2014)
21. Cianci, C.M., Raemy, X., Pugh, J., Martinoli, A.: Communication in a swarm of miniature robots: the e-puck as an educational tool for swarm robotics. In: International Workshop on Swarm Robotics, pp. 103–115. Springer (2006)
22. Collett, T.H., MacDonald, B.A., Gerkey, B.P.: Player 2.0: toward a practical robot programming framework. In: Proceedings of the Australasian Conference on Robotics and Automation (ACRA 2005), p. 145 (2005)
23. Cote, C., Brosseau, Y., Letourneau, D., Raïevsky, C., Michaud, F.: Robotic software integration using marie. Int. J. Adv. Robot. Syst. **3**(1), 55–60 (2006)
24. Darwin, C., Beer, G.: The origin of species. Dent (1951)
25. De Rosa, M., Goldstein, S., Lee, P., Campbell, J., Pillai, P.: Scalable shape sculpting via hole motion: motion planning in lattice-constrained modular robots. In: Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006, pp. 1462–1468. IEEE (2006)

26. Di Mario, E., Martinoli, A.: Distributed particle swarm optimization for limited-time adaptation with real robots. Robotica **32**(02), 193–208 (2014)
27. Einhorn, E., Langner, T., Stricker, R., Martin, C., Gross, H.M.: Mira-middleware for robotic applications. In: 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 2591–2598. IEEE (2012)
28. Elkady, A., Sobh, T.: Robotics middleware: a comprehensive literature survey and attribute-based bibliography. J. Robot. 2012 (2012)
29. Engelberger, J.F.: Robotics in Practice: Management and Applications of Industrial Robots. Springer Science & Business Media (2012)
30. Farinelli, A., Iocchi, L., Nardi, D.: Multirobot systems: a classification focused on coordination. IEEE Trans. Syst. Man Cybern. Part B (Cybernetics) **34**(5), 2015–2028 (2004)
31. Gerkey, B., Vaughan, R.T., Howard, A.: The player/stage project: tools for multi-robot and distributed sensor systems. In: Proceedings of the 11th International Conference on Advanced Robotics, vol. 1, pp. 317–323 (2003)
32. Gerkey, B.P., Vaughan, R.T., Stoy, K., Howard, A., Sukhatme, G.S., Mataric, M.J.: Most valuable player: a robot device server for distributed control. In: 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2001. Proceedings, vol. 3, pp. 1226–1231. IEEE (2001)
33. Gouaillier, D., Hugel, V., Blazevic, P., Kilner, C., Monceaux, J., Lafourcade, P., Marnier, B., Serre, J., Maisonnier, B.: The nao humanoid: a combination of performance and affordability. CoRR abs/08073223 (2008)
34. Grieco, L.A., Rizzo, A., Colucci, S., Sicari, S., Piro, G., Di Paola, D., Boggia, G.: Iot-aided robotics applications: technological implications, target domains and open issues. Comput. Commun. **54**, 32–47 (2014)
35. Gummadi, R., Gnawali, O., Govindan, R.: Macro-programming wireless sensor networks using kairos. In: International Conference on Distributed Computing in Sensor Systems, pp. 126–140. Springer (2005)
36. Habibi, G., Xie, W., Jellins, M., McLurkin, J.: Distributed path planning for collective transport using homogeneous multi-robot systems. In: Distributed Autonomous Robotic Systems, pp. 151–164. Springer (2016)
37. Heckel, F., Blakely, T., Dixon, M., Wilson, C., Smart, W.D.: The wurde robotics middleware and ride multirobot tele-operation interface. In: Proceedings of the 21st National Conference on Artificial Intelligence (AAAI06) (2006)
38. Howard, A., Parker, L.E., Sukhatme, G.S.: Experiments with a large heterogeneous mobile robot team: exploration, mapping, deployment and detection. Int. J. Robot. Res. **25**(5–6), 431–447 (2006)
39. Hu, G., Tay, W.P., Wen, Y.: Cloud robotics: architecture, challenges and applications. IEEE Netw. **26**(3), 21–28 (2012)
40. Jackson, J.: Microsoft robotics studio: a technical introduction. IEEE Robot. Autom. Mag. **14**(4), 82–87 (2007)
41. Jang, C., Lee, S.I., Jung, S.W., Song, B., Kim, R., Kim, S., Lee, C.H.: Opros: a new component-based robot software platform. ETRI J. **32**(5), 646–656 (2010)
42. Jia, S., Takase, K.: Network distributed monitoring system based on robot technology middleware. Int. J. Adv. Robot. Syst. **4**(1), 69–72 (2007)
43. Jia, S., Hada, Y., Gakuhari, H., Takase, K., Ohnishi, T., Nakamoto, H.: Intelligent home service robotic system based on robot technology middleware. In: 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4478–4483. IEEE (2006)
44. Jiang, S., Cao, J., Liu, Y., Chen, J., Liu, X.: Programming large-scale multi-robot system with timing constraints. In: 2016 25th International Conference on Computer Communication and Networks (ICCCN), pp. 1–9. IEEE (2016a)
45. Jiang, S., Liang, J., Cao, J., Liu, R.: An ensemble-level programming model with real-time support for multi-robot systems. In: 2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops), pp. 1–3. IEEE (2016)

46. Kernbach, S., Thenius, R., Kernbach, O., Schmickl, T.: Re-embodiment of honeybee aggregation behavior in an artificial micro-robotic system. Adapt. Behav. **17**(3), 237–259 (2009)
47. Kim, S., Laschi, C., Trimmer, B.: Soft robotics: a bioinspired evolution in robotics. Trends Biotechnol. **31**(5), 287–294 (2013)
48. Kramer, J., Scheutz, M.: Development environments for autonomous mobile robots: a survey. Auton. Robots **22**(2), 101–132 (2007)
49. Liang, J., Cao, J., Liu, R., Li, T.: Distributed intelligent mems: a survey and a real-time programming framework. ACM Comput. Surv. (CSUR) **49**(1), 20 (2016)
50. Lima, P.U., Custodio, L.M.: Multi-robot systems. In: Innovations in Robot Mobility and Control, pp. 1–64. Springer (2005)
51. Lopes, Y.K., Leal, A.B., Dodd, T.J., Groß, R.: Application of supervisory control theory to swarms of e-puck and kilobot robots. In: International Conference on Swarm Intelligence, pp. 62–73. Springer (2014)
52. Magnenat, S., Longchamp, V., Mondada, F.: Aseba, an event-based middleware for distributed robot control. In: Workshops and Tutorials CD IEEE/RSJ 2007 International Conference on Intelligent Robots and Systems, LSRO-CONF-2007-016. IEEE Press (2007)
53. Magnenat, S., Noris, B., Mondada, F.: Aseba-challenge: an open-source multiplayer introduction to mobile robots programming. In: Fun and Games, pp. 65–74. Springer (2008a)
54. Magnenat, S., Rétornaz, P., Noris, B., Mondada, F.: Scripting the swarm: event-based control of microcontroller-based robots. In: SIMPAR 2008 Workshop Proceedings, LSRO-CONF-2008-057 (2008b)
55. Magnenat, S., Rétornaz, P., Bonani, M., Longchamp, V., Mondada, F.: Aseba: a modular architecture for event-based control of complex robots. IEEE/ASME Trans. Mechatron. **16**(2), 321–329 (2011)
56. Makarenko, A., Brooks, A., Kaupp, T.: Orca: components for robotics. In: International Conference on Intelligent Robots and Systems (IROS), pp. 163–168 (2006)
57. Manso, L., Bachiller, P., Bustos, P., Núñez, P., Cintas, R., Calderita, L.: Robocomp: a tool-based robotics framework. In: International Conference on Simulation, Modeling, and Programming for Autonomous Robots, pp. 251–262. Springer (2010)
58. Mataric, M.J.: Interaction and intelligent behavior. Technical report, DTIC Document (1994)
59. McLurkin, J., Smith, J.: Distributed algorithms for dispersion in indoor environments using a swarm of autonomous mobile robots. In: In 7th International Symposium on Distributed Autonomous Robotic Systems (DARS). Citeseer (2004)
60. McLurkin, J., Smith, J., Frankel, J., Sotkowitz, D., Blau, D., Schmidt, B.: Speaking swarmish: Human-robot interface design for large swarms of autonomous mobile robots. In: AAAI Spring Symposium: To Boldly Go Where No Human-Robot Team Has Gone Before, pp. 72–75 (2006)
61. McLurkin, J., Lynch, A.J., Rixner, S., Barr, T.W., Chou, A., Foster, K., Bilstein, S.: A low-cost multi-robot system for research, teaching, and outreach. In: Distributed Autonomous Robotic Systems, pp. 597–609. Springer (2013)
62. McLurkin, J., Rykowski, J., John, M., Kaseman, Q., Lynch, A.J.: Using multi-robot systems for engineering education: teaching and outreach with large numbers of an advanced, low-cost robot. IEEE Trans. Educ. **56**(1), 24–33 (2013)
63. McLurkin, J., McMullen, A., Robbins, N., Habibi, G., Becker, A., Chou, A., Li, H., John, M., Okeke, N., Rykowski, J., et al.: A robot system design for low-cost multi-robot manipulation. In: 2014 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 912–918. IEEE (2014)
64. Michael, N., Fink, J., Kumar, V.: Experimental testbed for large multirobot teams. IEEE Robot. Autom. Mag. **15**(1), 53–61 (2008)
65. Mohamed, N., Al-Jaroodi, J., Jawhar, I.: Middleware for robotics: a survey. In: 2008 IEEE Conference on Robotics Automation and Mechatronics, pp. 736–742. IEEE (2008)
66. Mohamed, N., Al-Jaroodi, J., Jawhar, I.: A review of middleware for networked robots. Int. J. Comput. Sci. Netw. Secur. **9**(5), 139–148 (2009)
67. Mondada, F., Franzi, E., Guignard, A.: The development of khepera. In: Experiments with the Mini-Robot Khepera, Proceedings of the First International Khepera Workshop, LSRO-CONF-2006-060, pp. 7–14 (1999)

68. Mondada, F., Bonani, M., Raemy, X., Pugh, J., Cianci, C., Klaptocz, A., Magnenat, S., Zufferey, J.C., Floreano, D., Martinoli, A.: The e-puck, a robot designed for education in engineering. In: Proceedings of the 9th Conference on Autonomous Robot Systems and Competitions, IPCB: Instituto Politécnico de Castelo Branco, vol. 1, pp. 59–65 (2009)
69. Mottola, L., Picco, G.P.: Programming wireless sensor networks: fundamental concepts and state of the art. ACM Comput. Surv. (CSUR) **43**(3), 19 (2011)
70. Owens, G., Granader, Y., Humphrey, A., Baron-Cohen, S.: Lego® therapy and the social use of language programme: An evaluation of two social skills interventions for children with high functioning autism and asperger syndrome. J. Autism Dev. Disord. **38**(10), 1944–1957 (2008)
71. Parker, L.E.: Current state of the art in distributed autonomous mobile robotics. In: Distributed Autonomous Robotic Systems 4. Springer, pp. 3–12 (2000)
72. Prencipe, G., Santoro, N.: Distributed algorithms for autonomous mobile robots. In: Fourth IFIP International Conference on Theoretical Computer Science-TCS 2006, pp. 47–62. Springer (2006)
73. Pugh, J., Raemy, X., Favre, C., Falconi, R., Martinoli, A.: A fast onboard relative positioning module for multirobot systems. IEEE/ASME Trans. Mechatron. **14**(2), 151–162 (2009)
74. Quigley, M., Conley, K., Gerkey, B., Faust, J., Foote, T., Leibs, J., Wheeler, R., Ng, A.Y.: Ros: an open-source robot operating system. In: ICRA Workshop on Open Source Software, Kobe, Japan, vol. 3, p. 5 (2009)
75. Rogers III, J.G., Trevor, A.J., Nieto-Granda, C., Cunningham, A., Paluri, M., Michael, N., Dellaert, F., Christensen, H.I., Kumar, V.: Effects of sensory precision on mobile robot localization and mapping. In: Experimental Robotics, pp. 433–446. Springer (2014)
76. Rubenstein, M., Shen, W.M.: Automatic scalable size selection for the shape of a distributed robotic collective. In: 2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), pp. 508–513. IEEE (2010)
77. Rubenstein, M., Ahler, C., Nagpal, R.: Kilobot: a low cost scalable robot system for collective behaviors. In: 2012 IEEE International Conference on Robotics and Automation (ICRA), pp. 3293–3298. IEEE (2012)
78. Rubenstein, M., Cabrera, A., Werfel, J., Habibi, G., McLurkin, J., Nagpal, R.: Collective transport of complex objects by simple robots: theory and experiments. In: Proceedings of the 2013 International Conference on Autonomous Agents and Multi-agent Systems, International Foundation for Autonomous Agents and Multiagent Systems, pp. 47–54 (2013)
79. Rubenstein, M., Ahler, C., Hoff, N., Cabrera, A., Nagpal, R.: Kilobot: a low cost robot with scalable operations designed for collective behaviors. Robot. Auton. Syst. **62**(7), 966–975 (2014)
80. Rubenstein, M., Cornejo, A., Nagpal, R.: Programmable self-assembly in a thousand-robot swarm. Science **345**(6198), 795–799 (2014)
81. Saeedi, S., Trentini, M., Seto, M., Li, H.: Multiple-robot simultaneous localization and mapping: a review. J. Field Robot. **33**(1), 3–46 (2016)
82. Sapaty, P.: Military robotics: latest trends and spatial grasp solutions. Int. J. Adv. Res. Artif. Intell. **4**(4), 9–18 (2015)
83. Sartoretti, G., Hongler, M.O., de Oliveira, M.E., Mondada, F.: Decentralized self-selection of swarm trajectories: from dynamical systems theory to robotic implementation. Swarm Intell. **8**(4), 329–351 (2014)
84. Schlegel, C., Worz, R.: Interfacing different layers of a multilayer architecture for sensorimotor systems using the object-oriented framework smartsoft. In: 1999 Third European Workshop on Advanced Mobile Robots, 1999 (Eurobot'99), pp. 195–202. IEEE (1999)
85. Siciliano, B., Khatib, O.: Springer Handbook of Robotics. Springer Science & Business Media (2008)
86. Smart, W.D.: Is a common middleware for robotics possible? In: Proceedings of the IROS 2007 Workshop on Measures and Procedures for the Evaluation of Robot Architectures and Middleware. Citeseer, vol. 1 (2007)
87. Soares, J.M., Aguiar, A.P., Pascoal, A.M., Martinoli, A.: A graph-based formation algorithm for odor plume tracing. In: Distributed Autonomous Robotic Systems, pp. 255–269. Springer (2016)

88. Soares, J.M., Navarro, I., Martinoli, A.: The khepera iv mobile robot: performance evaluation, sensory data and software toolbox. In: Robot 2015: Second Iberian Robotics Conference, pp. 767–781. Springer (2016)
89. Stampfer, D., Lotz, A., Lutz, M., Schlegel, C.: The smartmdsd toolchain: an integrated mdsd workflow and integrated development environment (ide) for robotics software. J. Softw. Eng. Robot. **7**(1), 3–19 (2016)
90. Stoy, K., Nagpal, R.: Self-repair through scale independent self-reconfiguration. In: 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2004 (IROS 2004). Proceedings, vol. 2, pp. 2062–2067. IEEE (2004)
91. Tsui, K.M., Yanco, H.A.: Assistive, rehabilitation, and surgical robots from the perspective of medical and healthcare professionals. In: AAAI 2007 Workshop on Human Implications of Human-Robot Interaction, Technical Report WS-07-07 Papers from the AAAI 2007 Workshop on Human Implications of HRI (2007)
92. Utz, H., Sablatnog, S., Enderle, S., Kraetzschmar, G.: Miro-middleware for mobile robot applications. IEEE Trans. Robot. Autom. **18**(4), 493–497 (2002)
93. Volpe, R., Nesnas, I., Estlin, T., Mutz, D., Petras, R., Das, H.: The CLARAty architecture for robotic autonomy. In: Aerospace Conference, 2001, IEEE Proceedings, vol. 1, pp. 1–121. IEEE (2001)
94. Wang, M.M., Cao, J.N., Li, J., Dasi, S.K.: Middleware for wireless sensor networks: a survey. J. Comput. Sci. Technol. **23**(3), 305–326 (2008)
95. Whittier, L.E., Robinson, M.: Teaching evolution to non-english proficient students by using lego robotics. Am. Second. Educ. 19–28 (2007)
96. Wurman, P.R., D'Andrea, R., Mountz, M.: Coordinating hundreds of cooperative, autonomous vehicles in warehouses. AI Mag. **29**(1), 9 (2008)
97. Yan, Z., Jouandeau, N., Cherif, A.A.: A survey and analysis of multi-robot coordination. Int. J. Adv. Robot. Syst. **10** (2013)

Part VII
Interference Mitigation, Radiation Control, and Encryption

Interference Mitigation Techniques in Wireless Body Area Networks

Mohamad Jaafar Ali, Hassine Moungla, Mohamed Younis and Ahmed Mehaoua

Abstract A wireless body area network (*WBAN*) consists of a single coordinator and multiple low-power sensors to monitor the biological signals and functions of the human body. Due to the highly mobile and social nature of *WBAN*s, the performance of the individual *WBAN*s could be degraded due to the adverse co-channel interference that results from their uncoordinated coexistence with other nearby wireless networks. Like other wireless networks, energy and spectrum in *WBAN*s are both scarce resources. In fact, the limited spectrum resource is the cause of competition among *WBAN*s and leads to interference. Dynamic and efficacious energy and interference management taking into account the channel state and traffics are both crucial and substantial to further upgrade the capacity and enhance the goodness of user experience. In addition to the co-channel interference, the individual communication links between the body-mounted sensors and the *WBAN*'s coordinator experience high signal attenuation and distortion due to the nonhomogeneous nature of the human body tissue. In this chapter, the issues related to the coexistence among *WBAN*s and between *WBAN*s and other wireless networks will be analyzed. A comparative

M. J. Ali · H. Moungla · A. Mehaoua
LIPADE, University of Paris Descartes, Sorbonne Paris Cité,
Paris, France
e-mail: mohamadjali84@gmail.com; mali@silicom.fr

H. Moungla
e-mail: hassine.moungla@parisdescartes.fr

A. Mehaoua
e-mail: ahmed.mehaoua@parisdescartes.fr

M. J. Ali
SILICOM Paris, Parc Ariane, 15 Boulevard des Chénes, 78280 Guyancourt, France

H. Moungla
UMR 5157, CNRS, Institute Mines-Telecom, Télécom SudParis, Nano-Innov
CEA Saclay, France

M. Younis (✉)
Department of Computer Science and Electrical Engineering, University of Maryland,
Baltimore County, USA
e-mail: younis@umbc.edu

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_19

review of the radio co-channel interference mitigation and avoidance techniques in the literature will be provided. Furthermore, we show that the existing solutions fall short from achieving satisfactory performance, and outline open problems that warrant more investigation by the research community.

1 Wireless Body Area Networks Overview

The recent technological advances in wireless communication and microelectronics have enabled the development of low-power, intelligent, miniaturized sensor nodes that can be implanted in or attached to the human body. Inter-networking these devices is referred to as a *WBAN* and is revolutionizing remote health monitoring and telemedicine in general. In essence, a *WBAN* is a wireless short-range communication network that consists of a single coordinator and a finite number of low-power wireless sensor devices. These sensors enable continual monitoring of the physiological state of the body in stationary or mobility scenarios. The coordinator collects the measurements of the individual sensors and sends them to a gateway that in turn delivers the received data to a remote emergency monitoring station (command center) using the Internet or the cellular telecommunication infrastructure [1, 2]. Basically, the *WBAN* sensors monitor vital signs like blood pressure, sugar level, body temperature, oxygen saturation, CO_2 concentration, respiration rate, and electromyography. In addition, *WBAN* sensors may observe the heart (electrocardiography) and the brain (electroencephalographs) electrical activities as well as providing real-time feedback to the medical personnel. The aim of *WBAN*s is to simplify and improve speed, accuracy, and reliability of communication of sensors and/or actuators within, on, and in the immediate proximity of a human body. *WBAN*s can provide real-time patient monitoring and serve in various applications such as ubiquitous health care, telemedicine, entertainment, sports, and military [3]. However, the *IEEE 802.15.6* standard [4] classifies these applications into medical and nonmedical applications as shown in Table 1. A comparison of the characteristics of sample in-body and on-body sensors is shown in Table 2 [5]. Table 3 shows the list of SI units that we used throughout this chapter.

1.1 Classification of WBANs

In this section, we provide a brief overview of the various *WBAN* characteristics that affect its design and operation. Specifically, we classify *WBAN*s based on the types of nodes, number of nodes, network topology, and the communication architecture. *WBAN* nodes are independent devices that can differ in their functionality, implementation, and role. The functionality classification is as follows:

Table 1 *WBAN* applications

WBAN Applications			
Nonmedical	Medical		
	Remote control	Implant	Wearable
Entertainment applications	Patient monitoring	Cancer detection	Sleep staging
Real-time streaming	Telemedicine systems	Cardiovascular diseases	Wearable health monitoring
Emergency	Ambient-assisted living		Asthma
			Aiding professional and amature sport training
			Assessing soldier fatigue and battle readiness

Table 2 Characteristics of in-body and on-body applications [5]

Application type	Sensor	DR	DC	PC	QoS	PV
In-body	Glucose sensor	Few Kbps	<1%	Extremely	Yes	High
In-body	Pacemaker	Few Kbps	<1%	Low	Yes	High
In-body	Endoscope capsule	> Mbps	<50%	Low	Yes	Med
On-body medical	ECG	3 Kbps	<10%	Low	Yes	High
On-body medical	SpO2	32 Kbps	<1%	Low	Yes	High
On-body medical	Blood pressure	<10 bps	<1%	High	Yes	Med
On-body nonmedical	Music for headsets	1.4 Mbps	High	High	Yes	Low
On-body nonmedical	Forgotten things monitor	256 Kbps	Med	Low	No	Low
On-body nonmedical	Social networking	<200 Kbps	<1%	Low	No	High

Table 3 SI units

Symbol	Notation	Symbol	Notation
W	Watt	g	Gram
dB	Decibel	Hz	Hertz
m	Meter	mm	Millimeter
Kw	Kilowatt	Kg	Kilogram
KHz	Kilohertz	MHz	Megahertz
GHz	Gigahertz	mAh	Milliamp hour
dBm	Decibel-milliwatts	mW	Milliwatt
Kbp/s	Kilobit per second	Mb/s	Megabit per second

- *Sensors* are implanted inside or distributed on the human body to measure certain vital parameters such as blood pressure, blood glucose, etc.
- *Actuators* respond upon receiving commands from the coordinators. For example, an actuator pumps insulin into the bloodstream when a sensor reports a high percentage of glucose in the blood.
- *Personal Devices* (*PDs*) are responsible for gathering the information received from sensors, providing commands to actuators and interfacing the *WBAN* to external entities, e.g., hospital.

The *IEEE 802.15.6* standard [4] classifies the *WBAN* nodes according to their deployment as follows:

- An *implant node* (*in-body*) is placed either just underneath the skin or inside the human body as illustrated in Fig. 1.
- A *body surface node* (*on-body*) is usually at most 2 cm away from the surface of the human body as illustrated in Fig. 2.
- An *external node* is placed a few centimeters up to 5 m above the surface of the human body.

Based on their role, nodes can be classified as follows:

- A *coordinator* is a gateway or PDA-like node which connects the *WBAN* to the external entities.
- An *end node* is limited to performing application tasks and is not able to relay messages from other nodes.
- A *relay* is an intermediate node capable of sensing data and forwarding data from other nodes to the coordinator.

Table 4 provides high-level summary of node classification.

As per *IEEE 802.15.6* standard [4], the number of nodes in a typical *WBAN* network may range from 6 up to 256 within a 6 m $\times$ 6m $\times$ 6 m space. A single *WBAN* may involve a single coordinator and up to 64 nodes. Since 2 to 4 *WBAN*s may coexist on the same person (per $1m^2$), a maximum of 256 nodes may exist per person. The *IEEE 802.15.6* working group has considered *WBAN*s to operate in either a one-hop

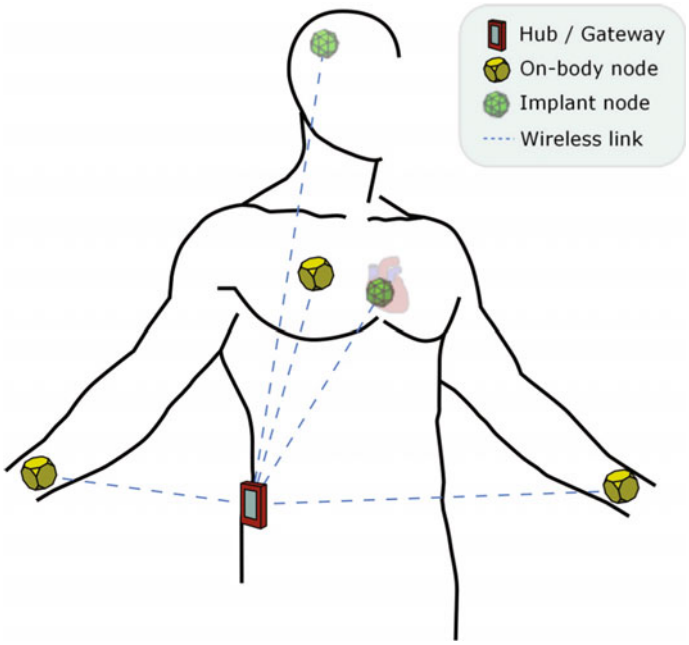


Fig. 1 A *WBAN* network illustrating a gateway (a *WBAN*'s coordinator), implant and on-body sensor nodes, and communication links among them [6]

or two-hop topology. In the one-hop star topology, two possible transmissions may exist. A transmission may be from the device to the coordinator and the other way around. In addition, two modes of communication may exist in the star topology, namely, beacon mode and non-beacon mode. In the beacon mode, the *WBAN* coordinator transmits beacons periodically to define the boundaries of a superframe and consequently enables nodes to synchronize their clocks. Basically, a superframe is defined as a fixed-length period of time and further divided into two other parts, active and inactive. The active part is divided into a fixed number of short periods of time, each called time slot, which are dedicated for sensors transmissions. While, the inactive part is also composed of fixed-length period of time and only dedicated for coordinator's transmission. In the non-beacon mode, a *WBAN* node can transmit data to the coordinator using CSMA/CA and can poll the coordinator to receive data. Meanwhile, in a two-hop topology, nodes are connected to the coordinator via other intermediate nodes called relays. Table 5 provides a comparison that shows the two-hop transmission has a higher delay and lower transmission power compared to the one-hop star topology and involves overheads along with its network operation as well as high complexity. More specifically, using relays lowers the concentration of the transmission power between the source and the destination. The latest version of the IEEE standard proposed for *WBAN*s [4] supports only two hops in IEEE *WBAN* standards compliant communication [7].

Fig. 2 A male subject wearing two body surface sensor nodes, the first sensor at the right wrist and the second sensor at the upper left arm [6]

Table 4 Node classification according to their functionality, deployment, and role

Node classification		
Functionality	Deployment	Role
Sensors (implanted, attached)	Implant node	Coordinator, e.g., gateway
Actuators	Body surface nodes	End node
Personal devices, e.g., sink/coordinator	External node	Relay

Table 5 Comparison of one-hop star network and two-hop network [8]

Comparison criteria	Star networks	Two-hop networks
Energy consumption	For nodes in close proximity to the PD, the power used to transmit to the PD will be low. The nodes further away, however, will consistently require more power to be able to transmit information	The nodes that are closest to the PD consume more energy as they will have to forward not only their own information but also information from other nodes
Transmission delay	The star network presents the least possible delay for transmitting data from any sensor to the PD, as there is only a single hop	The data of some nodes could experience delivery delay due to buffering at relays
Interference	Sensors that are farther away from the PD require transmission with a higher power, increasing the amount of interference	Since each node is only transmitting to its neighbors, the energy of transmission is kept low and hence the effects of interference stay limited
Node Failure and mobility	Only the failed node will be affected and the rest of the network can be performed as needed	The part of the network that involves the failed node has to be reconfigured. Overheads are involved

1.2 Communication Architecture of WBANs

The communication architecture of *WBAN*s consists of three tiers as illustrated in Fig. 3.

- ***Tier-1***: Intra-*WBAN* communication—This tier reflects the interaction of the *WBAN* nodes (sensors and coordinator). Sensors forward their measurements to the coordinator. The coordinator processes the collected physiological data and then transmits to an access point in Tier-2.
- ***Tier-2***: Inter-*WBAN* communication—This communication tier is between the coordinator and one or multiple access points. Tier-2 communication aims to interconnect *WBAN*s with various networks (cellular, etc.).
- ***Tier-3***: Caregiver—This tier reflects medical facility, nurses, doctors, and family members, and can be viewed as remote command centers.

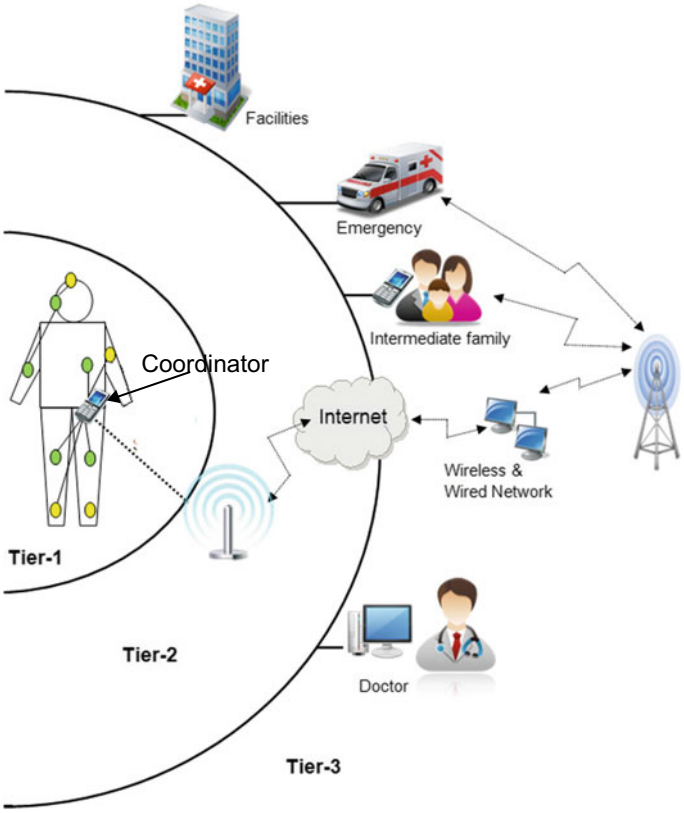


Fig. 3 Communication architecture of *WBAN*s

2 Intra-*WBAN* Communication

Though this chapter focuses mainly on interference mitigation techniques among *WBAN*s, it also discusses the intra-*WBAN* communication in details. Recently, the *IEEE 802.15.6* task group (TG6) has released the IEEE standard [4] for short-range wireless communications on or inside the human body. The standard has defined PHY and MAC protocols for *WBAN*s and whose functional requirements are summarized in Table 6. We further highlight the primary requirements and design considerations of wireless communication technologies which can be applied in *WBAN*s as follows.

- **Data rate**: to support variations among the *WBAN* applications, data rates should be ranging from 10 kbit/s to 10 Mbit/s. However, the bit error rate (BER) determines the reliability of the data transmission and is dependent on the criticality of the data, i.e., high data rate transmission requires low BER, e.g., 10^{-10}, and vice versa.
- **Transmission power**: *WBAN* sensors may transmit up to 1 mW (0 dBm) which complies with the specific absorption rate (SAR) 1.6 W/Kg in 1 g of the human

Table 6 *IEEE 802.15.6 WBAN* requirements [4]

Topology	One-hop star, Two-hop star, bidirectional link
Setup time	Insertion/removal <3 s
Devices number	Typically 6, up to 256
Data rate	10 Kb/s–10 Mb/s
Range	3 m with low data rate under IEEE Channel Mode
PER	<10% with a link success probability of 95% overall channel conditions
Latency	<125 ms (medical), <250 ms (nonmedical)
Jitter	<50 ms
Reliability	<1 s for alarm; < 10 ms for applications with feedback
Power consumption	>1 year (1% LDC + 500 mAh battery), >9 h (always ON + 50 mAh battery)
Intra-coexistence	10 *WBAN*s in a volume of 6 × 6 × 6 m
Inter-coexistence	Environment (WiFi, Bluetooth, etc,)

body tissue [9]. The battery lifetime of a *WBAN*'s sensor node should span several months or even years.

- **Mobility**: due to the postural body movements, e.g., sitting, walking, twisting, waving arms, etc., the *WBAN* may experience the signal shadowing and channel fading which detriment the reliability and QoS metrics such as packet delivery ratio, in-order-delivery, end-to-end delay (latency), transmission loss rate, delay, jitter, etc. Such reliability is necessary to protect the patient's life when a life-threatening event has not been detected. Thus, a highly reliable and energy-efficient data transfer with low delay is required to guarantee a successful data transmission for immediate access by health caregivers.
- **Configurable**: *WBAN* nodes should be configurable and capable of joining the *WBAN* and switching from a failed link to a backup link.
- **Coexistence**: *WBAN*s may interact and coexist with each other as well as with other wireless technologies. The co-channel interference may happen when different neighboring *WBAN*s and other wireless networks share similar channels at the same time which may lead to a dramatic increase in the level of interference. Therefore, a coexistence algorithm, that can be carried out independently or cooperatively, should guarantee a proper functionality of the different coexisting *WBAN*s.

Given the rigorous low-power consumption requirement in *WBAN* systems, the first wireless technology that could be used is the IEEE 802.15.1 standard (Bluetooth) [10] which was employed in many e-health and telemedicine applications. The properties of Bluetooth, namely, the bandwidth, lack of support of multi-hop communication, and long start-up time make it unsuitable for *WBAN* applications.

Bluetooth low power (BLE) [11] has been introduced as an amendment of the original Bluetooth and may be viewed as a better choice for *WBAN* applications. Basically, the lower power consumption can be achieved using low duty cycling; nonetheless, this exaggerated low-duty cycle mechanism makes BLE unsuitable for health monitoring applications as they need the high frequency of data transmissions.

On the one hand, the *IEEE 802.15.4* (ZigBee) standard [12] is widely employed in WSNs and has larger coverage area. However, the data rate of ZigBee is low which results in high latency of data delivery. Such a latency is unacceptable for sharing critical data in medical applications that report life-threatening conditions. In April 2010, the *IEEE 802.15.6* working group established the first draft of the communication standard of *WBAN*s that is optimized for low-power on-body/in-body nodes for various medical and nonmedical applications. The latest standardization of *WBAN*s, *IEEE 802.15.6* standard [13] aims to provide an international standard for low-power, short-range (within the human body), and extremely reliable wireless communication within the surrounding area of the human body, and support a vast range of data rates from 75.9 Kbps (narrowband) up to 15.6 Mbps ultra-wideband. Moreover, the standard utilizes different frequency bands as follows:

- The narrowband (NB) uses 400, 800, 900 MHz, and 2.3 and 2.4 GHz.
- The ultra-wideband (UWB) utilizes 3.1–11.2 GHz.
- The human body communication (HBC) utilizes the frequencies within the range of 10–50 MHz that cannot support high data rate transmissions, e.g., video or audio streaming.

The 2.4 GHz band is deemed by practitioners as the best option for the use by medical applications because of its ability against adjacent channel interference.

Due to the complex structure of the body shape and human tissue, a simple path loss model cannot be directly adopted for *WBAN* nodes because it does not consider the impact of the human tissue on radio signal propagation. Nodes in *WBAN*s can be implanted in or attached to the human body, which creates multiple transmission channels among them [14]. The communication authorities in the United States and other countries allocated the international unlicensed, ultra-low-power MICS band at 402–405 MHz with 300 KHz channels to support wireless communication with implanted medical devices. This allocation provides fast data transfer and long communication range as well as has conducive propagation characteristics for the transmission of radio signals through the human tissue and hence leads to better penetration compared to higher frequencies (10 dB for 10 mm tissue penetration) [15]. The channel models proposed by *IEEE 802.15.6* standard are shown in Table 7 [14].

Khan et al. [16] have shown that a log-normal distribution better suits long-term fading and that the Rician distribution best fits short-term fading for on-body channels at 2.45 GHz, 5.8 GHz, and 10 GHz. Smith et al. [6] have concluded that a log-normal distribution is the best fit model for small-scale fading in ultra-wideband communications. Meanwhile, a log-normal distribution or a gamma distribution suits small-scale fading in narrowband communications. Various models to describe channel fading

Table 7 Scenarios and descriptions of channel models in *IEEE 802.15.6* [14]

Scenario	Desc.	Frequency band	Channel model
S1	Implant to implant	402–405 MHz	CM1
S2	Implant to body surface	402–405 MHz	CM2
S3	Implant to external	402–405 MHz	CM2
S4	Body surface to body surface (LOS)	13.5, 50, 400, 600, 900 MHz, 2.4, 3.1–10.6 GHz	CM3
S5	Body surface to body surface (NLOS)	13.5, 50, 400, 600, 900 MHz, 2.4, 3.1–10.6 GHz	CM3
S6	Body surface to body surface (LOS)	13.5, 50, 400, 600, 900 MHz, 2.4, 3.1–10.6 GHz	CM4
S7	Body surface to external (LOS)	3.5, 50, 400, 600, 900 MHz, 2.4, 3.1–10.6 GHz	CM4

in *WBAN*s have been compared in [6, 17–22]. The most common distribution fit for *WBAN*s is log-normal, followed by Nakagami-m then Rician.

3 Radio Co-channel Interference

In hospitals and crowded public places, *WBAN*s may coexist with other wireless networks such as IEEE 802.11 (WiFi), *IEEE 802.15.4* (ZigBee), WSNs, IEEE 802.15.1 (Bluetooth), MANETs, cellular and other appliances that may share the same international unlicensed ISM band 2.4 GHz as *WBAN*s. An example of radio co-channel interference that can be experienced between two coexisting *WBAN*s and other wireless networks is shown in Fig. 4 [9]. Basically, the co-channel interference problem may exist among homogeneous *WBAN*s, e.g., a patient with a *WBAN* may be surrounded by other patients with *WBAN*s, which impose the interference on each other, i.e., *WBAN*-*WBAN* interference, as shown in Fig. 2. On the other hand, the interference problem may also be experienced across multiple different heterogeneous wireless networks, i.e., nodes from non-*WBAN*s wireless networks, e.g., WiFi, impose the interference on some nodes of the *WBAN*s, i.e., WiFi-*WBAN* interference, as shown in Fig. 5.

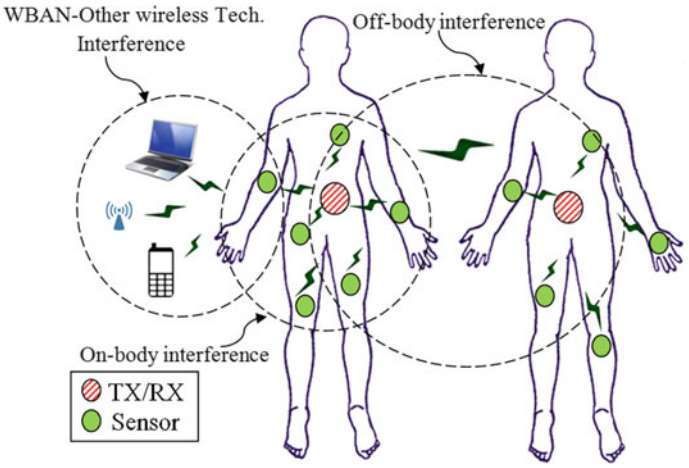


Fig. 4 Radio co-channel interference between *WBAN*s and other wireless networks

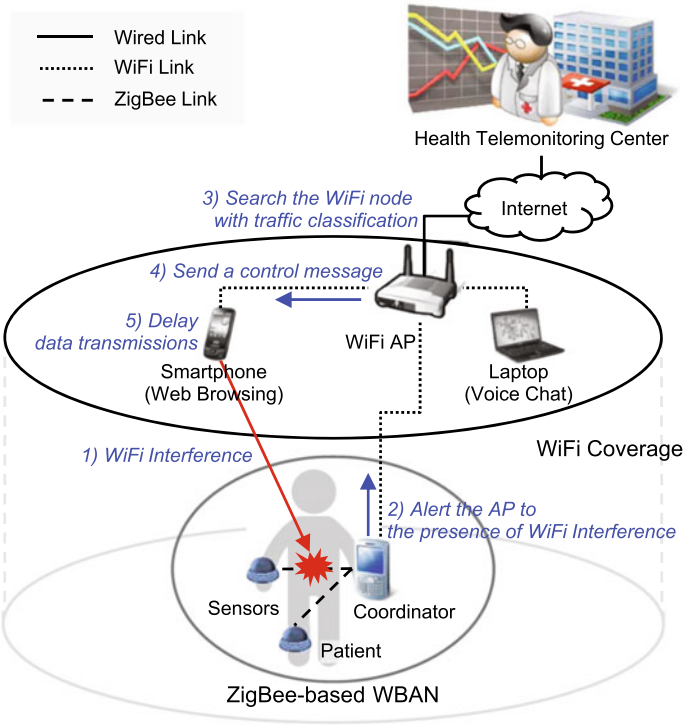


Fig. 5 Radio co-channel interference between one *WBAN* and one WiFi wireless network

3.1 Interference Between WBANs and Other Wireless Networks

Dealing with the co-channel interference problem within the same *WBAN* is easier than across networks due to many reasons. First, the use of the different MAC protocols among coexisting wireless networks may lead to medium access collisions, as these *WBAN*s share similar channels. Second, some wireless networks like WiFi may use larger bandwidth and packet size than those used by *WBAN*s, which will reserve the medium for the longer period of time that leads to unfair usage of the medium. Third, WiFi networks may use higher transmission power level than *WBAN*s and consequently they dominate the medium and hinder intra-*WBAN* communication. Accordingly, the co-channel interference will impact performance metrics including energy lifetime, data reliability, throughput, delay, success packet delivery ratio, etc.

The heterogeneity of wireless technologies used by other wireless networks may introduce new challenges for *WBAN* system designers. *WBAN*s could be subject to more frequent topology changes due to the human body mobility and move faster than the conventional WSNs. *WBAN*s move in a group-based rather than node-based manner as in MANETs. The nodes in *WBAN*s are deployed more densely in a very small area, while the locations of mobile stations in the cellular networks are spread over a wide area. Therefore, the interference mitigation schemes for WSNs, MANETs, and cellular networks do not suit *WBAN*s. Thus, dealing with radio co-channel interference among *WBAN*s and due to their coexistence with other wireless networks is an important problem that warrants special attention [9].

3.2 Radio Co-channel Interference Among WBANs

*WBAN*s are becoming pervasive. Their coexistence will become a serious issue in the upcoming years. In 2009, eleven million sensors were estimated in use. Such a number is predicted to reach 485 million by 2018 [9, 23, 24]. As pointed out in Sect. 2, the *IEEE 802.15.6* standard [13] requires the system to function properly within the transmission range of up to 3 m when up to 10 *WBAN*s are collocated. Moreover, the system should also be able to support up to 60 sensors in a 6 m^3 space (256 sensors in a 3 m^3). As per *IEEE 802.15.6* standard, the superframe is delimited by two beacons and composed of two successive frames: (i) active, that is dedicated for sensors, and (ii) inactive, that is designated for coordinators as illustrated in Fig. 6.

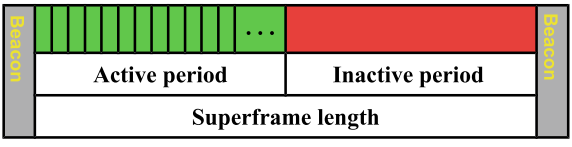


Fig. 6 *IEEE 802.15.6* superframe structure illustrating active and inactive periods [13]

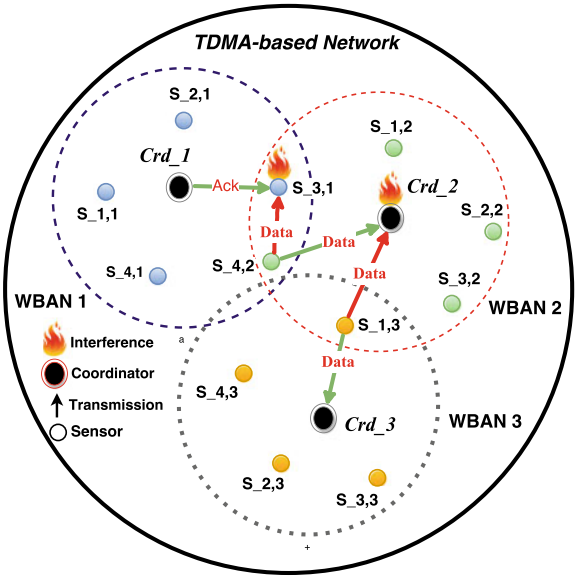


Fig. 7 Radio co-channel interference among three TDMA-based *WBAN*s

Due to the typical social interaction of people, individual *WBAN*s may move towards each other in crowded areas such as a hospital's lobby or public places. When these *WBAN*s operate on similar channels, the corresponding radio communication ranges, and the individual active periods of their superframes, i.e., TDMA- or CSMA/CA-based superframe may overlap with each other [25, 26]. Basically, if a large number of sensors of different *WBAN*s coexist in the close proximity of each other, there is a high probability that they access the same channel at the same time and their transmissions may cause medium access collision as illustrated in Fig. 7. Although *WBAN*s search for available channels, the interference occurs because of the smaller number of channels in the *IEEE 802.15.6* standard than the number of coexisting *WBAN*s [4, 12, 13]. Even when few *WBAN*s coexist, such interference may affect the communication links and may have an adverse impact on the reliability of *WBAN*s, by decreasing the SINR of the received signal resulting in extra packet losses and performance degradation. Where the signal-to-interference-plus-noise ratio (SINR) is defined as a quantity used to give theoretical upper bounds on channel capacity in wireless communication systems. Therefore, the co-channel interference mitigation is of the utmost importance to improve the reliability and the performance of the whole network. An example of radio co-channel interference that can be experienced among coexisting *WBAN*s is shown in Fig. 8 [9].

On the other hand, the resource-constrained nature of *WBAN* nodes in terms of limited power supply, i.e., small battery capacity, size, transmission range, and cost makes the application of advanced antenna and power control techniques used in other wireless networks unsuitable for *WBAN*s. The protocols proposed for these networks do not consider the special characteristics and stringent properties of *WBAN* systems

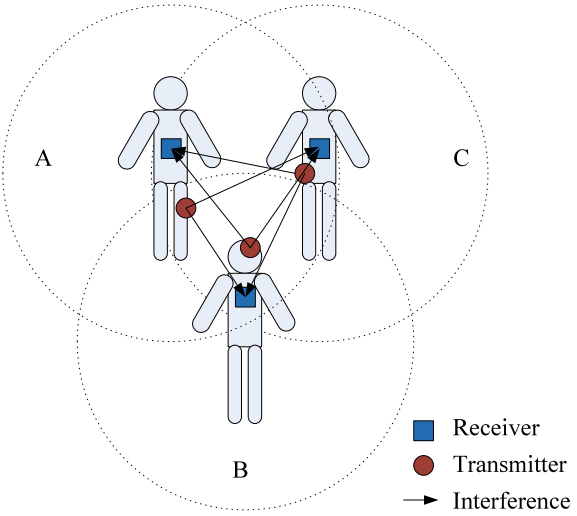


Fig. 8 Radio co-channel interference among coexisting *WBAN*s

[27–32]. For instance, power control mechanisms which proved their efficiency in cellular systems are not suitable for *WBAN*s since they require high transmission power, which requires larger sensors batteries. Another example, the simple design, and shape of sensor's antenna make signal processing very hard to be achieved in *WBAN*s because of the inhomogeneous nature of the human body characterized by high signal attenuation and distortion. Moreover, due to the highly mobile nature of *WBAN*s, e.g., walking, running, bending, etc, different *WBAN*s may change their position relative to each other. In addition to the body posture, e.g., movements of arms and legs, the individual sensors in the same *WBAN* may change their location relative to each other. Furthermore, the absence of coordination and negotiation as well as the synchronization among the coordinators of different *WBAN*s make the allocation of a centralized entity to manage *WBAN*s coexistence and mitigate the interference unsuitable for *WBAN*s [1, 2, 23, 33, 34]. In fact, this shortening is due to the fast-changing *WBAN*'s and the network's topologies and in consequence the high rate of update and control messages within the whole network.

The *IEEE 802.15.6* standard recommends the use of TDMA as an alternative solution to avoid co-channel interference within *WBAN*s. To this end, it proposes the following three mechanisms for inter-*WBAN* co-channel interference mitigation [2, 35]:

- **Beacon Shifting**: a *WBAN* coordinator may transmit its beacons at different time offsets relative to the start of the beacon periods by including a beacon shifting sequence field in its beacons. A coordinator should choose a beacon shifting sequence that is not being used by its neighbor coordinators to mitigate the potential interference.
- **Channel Hopping**: a coordinator may change its operating channel periodically by choosing a particular channel hopping sequence, which is not being used by any

neighbor coordinators. Moreover, the coordinator should hop to another channel after dwelling on the current channel for a fixed number of beacon periods and not in the middle of a superframe.
- **Interleaving**: a *WBAN* may share the same operating channel with one or more other *WBAN*s. With interleaving, the *WBAN*s' coordinators exchange information in order to prevent the active periods of their corresponding superframes from overlapping with each other.

3.3 Interference Mitigation Challenges

In this section, we provide a brief overview of the challenges in applying an interference mitigation without negatively impacting the performance of the individual *WBAN*s. An efficient interference mitigation technique should carefully balance between the excessive use of the scarce and limited resources in *WBAN*s and the desired requirements of a *WBAN* application.

3.3.1 Network Throughput

Network throughput is the most important metric to gauge the performance of a *WBAN* since it implicitly captures the packet delivery rate (PDR). When the throughput of the *WBAN* grows, the reliability of the communication in the whole *WBAN* increases, i.e., fewer packet losses. In healthcare applications, the *WBAN* sensors must ensure an efficient interference mitigation, consequently successful delivery of patient's vital data. Therefore, the interference mitigation technique must optimize the network throughput.

3.3.2 Energy Consumption

To ensure a long lifetime of *WBAN*, energy efficiency is always desired for their sensors because of the small size of sensors' batteries and the difficulty in recharging and replacing them. Due to co-channel interference, collisions may happen, which leads to increased packet retransmissions and consequently, the energy consumptions of *WBAN* sensors grow accordingly. In addition, *WBAN*s may experience severe interference from other wireless networks and hence they may transmit with higher power levels to compete for better signal-to-interference-plus-noise ratio. Therefore, a trade-off between WBAN performance and energy consumption is required to guarantee that the interference is efficiently reduced.

3.3.3 QoS and Reliability

*WBAN*s have specific QoS requirements which depend on the BER or the priority of the traffic. In crowded situations, e.g., in hospitals where there are adverse interference, *WBAN*s with high QoS constraints must have a high priority to access the channel because some patients may report vital data, e.g., heart disease data. Generally, reliability is related to packet delay and the probability of packet loss. We define the convergence time as the time needed for the interference mitigation technique to enable a *WBAN* to operate normally. Thus, the faster the interference mitigation converges, the more effective it is. In situations of high interference, the convergence time impacts the packet delay. In addition, the probability of packet loss specifies the range to which the packet drop rate impacts the reliability in terms of BER or PER of the *WBAN*.

3.3.4 Dynamic Environment

It is necessary for any interference mitigation technique to efficiently handle *WBAN*s coexistence in dynamic and mobile environments. Employing inter-*WBAN* negotiation as a means for limiting interference may result in long convergence time or delay. On the other hand, in the absence of negotiation, each *WBAN* tries to optimize its individual performance selfishly, converge rapidly, and learn about the network independently. Thus, it is desirable to strike a balance between these conflicting design choices.

3.3.5 Impact of Radio Frequency Radiation on the Human Body

As the sensors are implanted in or attached to the human body, the radio frequency radiation could affect the human body. Therefore, interference mitigation techniques with minimal adverse radiation are desirable to ensure the short-term and long-term safety of the human body.

4 Co-channel Interference Mitigation Techniques

Avoidance and mitigation of radio co-channel interference have been extensively researched in the wireless communication literature. Several studies have focused on the adverse effects of radio co-channel interference on the performance of *WBAN*s [36–40]. Published co-channel interference mitigation and avoidance techniques can be categorized as resource allocation, power control, link adaptation, and incorporation of multiple medium access arbitration mechanisms [38, 41–48]. The purpose of the co-channel mitigation and avoidance techniques is to ensure that *WBAN*s operate in a stable way even in the populated area, under high mobility conditions, and in

Table 8 Symbols and notations

Symbol	Notation	Symbol	Notation
Med	Medium	Dyna	Dynamic
MOB	Mobility	TPO	Topology
COP	Cooperation	DEL	Delay
TOF	Trade-off	DR	Data rate
SPR	Spatial reuse	CMX	Complexity
NEG	Negotiation	CHST	Channel status
CHP	Channel parameter	REL	Reliability
CHUT	Channel utilization	THR	Throughput
MAC	Medium access control	CEX	Coexistence
CNV	Convergence time	EC	Energy consumption
LCR	Level crossing rate	OP	Outage probability

situations of a high level of co-channel interference. Table 8 shows the list of symbols and the corresponding notations that we used in the balance of this subsection.

4.1 Resource Allocation

Countering interference by careful resource management and medium access scheduling is obviously a viable option. Resource allocation, e.g., channels and time, is an effective way for avoiding radio co-channel interference and medium access collision. Some approaches have pursued this methodology. We group the published work into three categories as we will discuss in the balance of this subsection.

4.1.1 Channel Assignment

Channel assignment deals with the allocation of channels to individual sensors, coordinators, or any combination of them. Once the channels are allocated, *WBAN* coordinators may then allow the individual sensors or another coordinator within the network to communicate via the available channels. The objective is to achieve maximum system spectral efficiency in bits/Hz by means of frequency reuse, but still, assure a certain grade of service by avoiding co-channel interference and adjacent channel interference among nearby *WBAN*s or wireless networks that share the bandwidth. The main problem in channel assignment solutions is the limited number of available channels, especially, when there is high density of coexisting *WBAN*s. Moreover, there is no accurate methodology to determinate the level of interference

based on SINR, RSSI, channel quality, etc. Channel assignment can be categorized into three approaches [49]:

- **Packet-based approach** assigns channels on a per-packet basis at a given node. Obviously, this approach could impose high overhead and lead to increased data losses given the frequent changes in the used channels. No wonder no published scheme pursued a per-packet channel assignment.
- **Link-based approach** assigns a channel for each link between two given nodes. All packets between the two nodes will be transmitted on the same channel.
- **Flow-based approach** where all packets belonging to a flow are sent on the same channel.

Few published protocols pursued *Link-based* channel assignment. For example, LAH [41] is based on adaptive channel hopping. Such channel hopping is decided according to the combination of a set of interference detection parameters (beacon delivery rate, RSSI, etc.). LAH does not need information exchange among *WBAN*s and is shown to improve the network throughput and lifetime. DRS [50] is a distributed dynamic resource allocation inter-*WBAN* interference mitigation scheme. In DRS, interference-free sensors from different *WBAN*s transmit on the same channel, while highly interfering sensors transmit using orthogonal channels in order to maximize the spatial reuse. In DRS, the coordinators need to exchange SINR information with each other. Moreover, the resource allocation performs better for a static than dynamic *WBAN* network topology, i.e., due to *WBAN* mobility. The approach of [51] is to minimize the impact of co-channel interference through dynamic channel hopping based on Latin rectangles [52]. Two schemes for channel allocation and medium access scheduling diminish the probability of inter-WBAN interference. The first scheme, namely, DAIL, assigns channel and time slot combinations that reduce the probability of medium access collision. No inter-WBAN coordination is needed in DAIL. Despite being very effective, DAIL involves overhead due to frequent channel hopping at the coordinator and sensors. The second scheme, namely, CHIM, takes advantage of the relatively lower density of collocated WBANs and pursues hopping among channels only when interference is detected in order to save power. Basically, CHIM provisions backup time slots for failed transmission. The backup time slots are scheduled in a way that is similar to DAIL. CHIM enables only a sensor that experiences collisions to hop to an alternative backup channel in its allocated backup time slot. Meanwhile, AIM is a *flow-based* approach. AIM [42] classifies the sensors transmissions according to the QoS, packet length, SINR, etc. AIM allocates an orthogonal channel to each sensor that has the highest priority and has not been scheduled yet. Since AIM considers sensor-level interference mitigation, it significantly reduces the number of assigned channels as well as achieves a higher throughput.

On the other hand, some approaches tried to avoid interference by assigning conflict-free channels. Obviously, this approach does not suit dynamic environments where *WBAN*s accidentally come in range of each other. Nonetheless, conflict-free channel assignment can fit scenarios where set of *WBAN*s that may coexist can be predicted beforehand. The popular methodology for channel assignment in this case

is to use graph coloring. Some approaches such as RIC [53] assume global control for assigning a channel to each *WBAN* using a lightweight random incomplete coloring algorithm. Although RIC allows for higher spatial reuse of channel allocation, it only considers channel assignment at the *WBAN* level rather than at the node level, and hence the spectrum is not optimally utilized. Meanwhile, GCS [54] is a hybrid scheme integrating graph coloring and cooperative scheduling for *WBAN*s. GCS pairs every two *WBAN*s into a cluster and uses cooperative scheduling among each *WBAN*s of the cluster to minimize interference and increase the spatial reuse. A graph coloring scheme is then applied for channel allocation for *WBAN*s, which adds no complexity to sensors as the coordinators only negotiate channel assignment. Although the algorithm improves the power saving of sensor nodes, it increases the time needed by sensors to complete their transmissions, which is undesirable in a *WBAN*.

4.1.2 Transmission Scheduling

Intuitively the medium may be shared on a time basis. Basically, data packet rescheduling is used to mitigate interference by assigning unused time slots. Some interference mitigation schemes pursued careful scheduling of sensor transmissions so that medium access collision could be avoided. Yan et al. [36] presented a QoS-driven transmission scheduling approach to limit the duration that a node in a *WBAN* has to be in active mode under time-varying traffic and channel conditions. The approach, which is named QSC, optimally assigns time slots for each sensor node according to the QoS requirement while minimizing their energy consumption. CWS [55] clusters sensors of different *WBAN*s into groups that avoid node-level interference. Then, CWS maps groups to the available time slots using the random coloring algorithm. CWS improves the system throughput and the network lifetime. Similarly, CSM [56] is a graph coloring-based scheduling method that avoids the inter-*WBAN* interference by assigning different time slots to adjacent *WBAN*s and by allocating more time slots to traffic-intensive *WBAN*s to increase the overall throughput. In CAG [57], different time slots are mapped to distinct colors and a color assignment is found for each node in the network. The *WBAN* coordinators exchange messages to achieve a conflict-free coloring in a distributed manner.

4.1.3 Combined Channel and Time Allocation

A number of approaches try to mitigate interference by considered channel and time allocation collectively. Basically, variations in channel assignment due to mobility scenarios within each *WBAN* as well as the movement of *WBAN*s towards each other are factored in when allocating time slots. Accordingly, Movassaghi et al. [58] proposed a distributed prediction-based inter-*WBAN* interference algorithm for channel

allocation. The algorithm, which is called CAS, allocated transmission time based on such prediction-based channel allocation in order to reduce the number of interfering sensors, extend *WBAN* lifetime, and improve the spatial reuse and throughput. Similarly, ACT [49] is an adaptive scheme to allocate channel and time for *WBAN*s in order to improve the throughput. ACT adaptively allocates channel and time simultaneously by considering the channel conditions and the density of *WBAN*s. Unlike CAS and ACT, JAD [9] is an adaptive scheme based on social interaction (JAD). By knowing the mobility pattern of *WBAN*s, JAD factors in traffic load, RSSI, and the density of sensors in a *WBAN* to optimize transmission time to minimize the interference and power consumption and to increase the throughput. An energy-efficient channel assignment is further proposed to reduce power consumption using transmit power control with simple channel prediction. *Overall, resource allocation protocols, when applied in WBANs, must take topology and link changes as well as the dynamic traffic into account. If carefully designed, these protocols may work efficiently under the high level of interference and mobility conditions. Nonetheless, they require a frequent exchange of information among WBANs and lead to a cost (e.g., energy, delay, etc.) in updating information. While, graph-based resource allocation protocols do not suit the dynamic environment and are unsuitable for a topology with high frequent changes, e.g., WBANs, because of the incurred cost due to update and message exchanges. In highly mobile and densely deployed WBANs, a graph-based resource allocation protocol may be not only inefficient but also detrimental for healthcare applications. In addition, the convergence time of the graph-based algorithm is an important issue that needs a careful attention in WBANs. Under high level of interference and mobility conditions, these protocols do not even support an acceptable level of QoS requirements to sensors.*

4.1.4 Summary of Resource Allocation

Channel and time resource allocation protocols may work efficiently in a dynamic environment and under high interference conditions, i.e., highly mobile and densely deployed *WBAN*s. These protocols may support an acceptable level of QoS requirements. Graph-based resource allocation protocols are unsuitable for a topology with high frequent changes, e.g., *WBAN*s, because of the incurred cost due to update and message exchanges. Therefore these protocols may be not only inefficient but also detrimental for healthcare applications in the dynamic environment. Moreover, they do not even support an acceptable level of QoS requirements to sensors. Table 9 provides a comparative summary of the different channel, time, and hybrid allocation interference mitigation proposals discussed in this subsection.

Table 9 Comparison of published resource allocation interference mitigation proposals for *WBAN*s. A star topology is deployed in the following proposals:

	EC	REL	THR	SPR	DEL	QoS	COP	MOB	CEX	CMX	CNV	MAC
JAD [9]	Low	High	High	High	Low	Yes	Yes	Yes	Yes	N/A	N/A	TDMA
ACT [49]	Low	High	High	N/A	N/A	Yes	No	Yes	Yes	N/A	N/A	TDMA
LAH [41]	High	Low	Low	Low	High	No	No	No	Yes	N/A	N/A	CSMA
DRS [50]	Med	Low	Low	High	High	No	Yes	No	Yes	N/A	N/A	TDMA
AIM [42]	Med	Med	High	Med	High	Yes	Yes	No	Yes	N/A	N/A	TDMA
RIC [53]	Low	N/A	High	High	N/A	N/A	No	Yes	Yes	Low	Fast	TDMA
GCS [54]	Low	N/A	High	High	N/A	N/A	No	Yes	Yes	Med	Slow	TDMA
QSC [36]	Low	Med	Med	N/A	Med	Yes	No	No	No	N/A	N/A	TDMA
CWS [55]	Low	N/A	High	High	N/A	N/A	Yes	No	Yes	Med	Slow	TDMA
CSM [56]	N/A	High	N/A	N/A	N/A	N/A	No	No	Yes	Low	Fast	TDMA
CAG [57]	Low	N/A	N/A	N/A	N/A	N/A	No	No	Yes	Med	Fast	TDMA
CAS [58]	Low	High	High	High	N/A	Yes	Yes	Yes	Yes	N/A	N/A	TDMA

4.2 Power Control

4.2.1 Link-State Based Power Control

One of the most design issues in *WBAN*s is the scarce energy resource of their sensors. The radio transceiver is considered as the most energy-consuming part in a sensor node. Thus, saving energy by adjusting transmission power is very crucial to extend the lifetime of the *WBAN*. Since channel conditions change frequently in a *WBAN* due to the body movements, the instantaneous channel state information is often unavailable to the transmitter. Moreover, the high path losses and attenuation of the wireless signals near the human body reduce the energy efficiency of *WBAN*s. In *WBAN*s, different factors such as fading, path loss and shadowing due to the human body tissue and mobility determine the link-state quality. However, the transmission power can be adaptively controlled based on its link state and hence, a trade-off exists between the energy consumption and link reliability. Many TPC protocols [59] have been employed in other wireless networks such as WSNs, WLANs, and cellular networks that used the lowest transmission power levels in order to achieve more energy efficiency, high link reliability, and reduced interference. Centralized TPC solutions proved their efficiency in wireless cellular networks and WSNs which have fewer resource constraints and more stable network topology than *WBAN*s. However, these solutions are unsuitable for dynamic and highly mobile *WBAN*s as each *WBAN* works independently [60, 61]. TPC protocols can be classified into three different mechanisms:

- **Connectivity-based mechanisms** consider the number of incoming neighbors (e.g., increasing node's transmission power adds more nodes in its communication range) to adapt the transmission power for each link, which is suitable for dense *WBAN*s, but unsuitable for *WBAN*s consisting of few sensor nodes [62].
- **Packet reception rate (PRR) based mechanisms** select the transmission power that ensures the PRR exceeds a predefined threshold. However, these mechanisms cannot immediately adjust the transmission power if the link state varies [43, 44].
- **RSSI-based mechanisms** that evaluate the link state based on the RSSI, where a higher RSSI indicates a better link quality. Some work such as the adaptive TPC mechanism of [63] predicts the suitable transmission power according to RSSI equations derived by empirical experiments rather than real-time RSSI measurements.

See et al. [43] and Ge et al. [44] conducted different experiments to capture the PRR corresponding to RSSI variation measured in static and dynamic body posture scenarios for healthcare applications at 2.48 GHz. A correlation between the path loss and the PRR was made via the probability distribution of the RSSI for a given transmit power. In addition, the optimal transmit power at the different locations of sensors was obtained in order to conserve the battery energy. Quwaider et al. [37] developed a dynamic body posture-based power control mechanism (DOI) based on RSSI. DOI provides optimal power assignments for the links among sensors of a *WBAN* while

Table 10 Comparison of published link-state based power control interference mitigation proposals for *WBAN*s

	THR	EC	REL	MOB	NEG	CHP	TOF
DOI [37]	Med	Med	Med	Yes	No	RSSI	No
DTP [64]	High	Low	High	Yes	No	RSSI	No
RTR [65]	Med	Low	Med	Yes	No	RSSI	Yes
LSE [66]	High	Med	High	Yes	No	RSSI	No
HOS [67]	High	Med	High	No	Yes	SINR	Yes
AGA [68]	Med	Low	Med	No	Yes	SINR	Yes

using the lowest power level to maintain high PDR. However, DOI predicts incorrect transmission power when the link state varies frequently. Whilst, Guan et al. [64] proposed a dynamic TPC DTP algorithm that achieves better energy saving and high link reliability in a mobile *WBAN*. Basically, DTP calculates the adjustment in transmission power according to the variation of the channel conditions.

In [65], Xiao et al. promote a real-time reactive scheme (RTR) that adjusts the transmission power according to the RSSI feedback from the receiver, under different mobility conditions. Similarly, the link-state estimation TPC protocol (LSE) [66] adapts the transmission power according to short-term and long-term link-state estimations. The short-term estimations were generated from several RSSI samples and the long-term estimations were generated through adjusting the RSSI threshold range according to variations in RSSI samples. RSSI variations are investigated according to the stationary state, posture change, and dynamic body motion of a patient. LSE is shown to achieve lower transmission power levels, lower packet loss, and RSSIs than those achieved in the other protocols.

On the other hand, HOS [67] is an opportunistic scheduling scheme that considers node-level interference mitigation and energy harvesting to extend the *WBAN*s energy lifetime. Low interfering sensor nodes transmit on the same channel while high interfering sensor nodes transmit using orthogonal channels. Sensor nodes harvest energy from the wireless of other nodes in the network. Thus, HOS uses the interference as a source of energy. Meantime, AGA [68] is a power allocation algorithm based on genetic algorithm (GA) to mitigate inter-*WBAN* interference while ensuring heterogeneous QoS guarantees. An optimization model using GA is utilized to minimize the transmission power of the sensor in a *WBAN*. However, AGA did not consider the real-world mobility which leads to a long convergence time. *Overall, link-state based power control protocols are suitable for highly mobile and densely network with WBANs. However, they do not suit dynamic channels environment because the link state varies very frequently.* Table 10 *provides a comparative summary of the different link-state based power control interference mitigation proposals discussed in this subsection.*

4.2.2 Game Theory

Game theory is a mathematical tool to study the interaction among several decision makers that may share a set of common conflicts or interests. Basically, the game model involves a set of players that select their actions in each round of the game to maximize some utility function reflecting their interest. Some examples of the utility in the context of *WBAN*s could be energy, throughput, delay, etc. The utility of the game usually fluctuates until stabilizing where all players get the best possible gain. Such stability stage is called Nash equilibrium (NE). When players unilaterally decide on their strategy, the game is called noncooperative. On the hand, the notion of a cooperative game reflects the fact that players collaboratively decide on the game rules. Game theory has been popular in the context of power control in *WBAN*s. It has been shown in [69, 70] that non-cooperative games are more appropriate for mitigating inter-*WBAN* interference since *WBAN*s operate independently; meanwhile pursuing cooperative games, i.e., message and information exchange among *WBAN*, increases the energy consumption. We review published techniques in the balance of this section.

Cooperative Games

Very few research studies have pursued the cooperative game theory approach for controlling transmission power in *WBAN*s. Gengfa et al. [71] proposed an inter-*WBAN* interference aware proactive power control algorithm (PAU) motivated by game theory that considers a limited cooperation among *WBAN*s. Each *WBAN* must exchange information about its current transmit power and channel gain with other interfering *WBAN*s. Although PAU has fast convergence time and low overhead, it is unsuitable for mobile *WBAN*s because it assumes the channel and interference gains stay fixed, and it does not consider the QoS requirements for sensors. Similarly, Wang et al. [70] proposed a distributed cooperative scheduling scheme (CSR) to reduce inter-*WBAN* interference and increase the throughput. CSR formulates single-*WBAN* scheduling as an assignment problem which has been solved by using horse racing scheduling algorithm. The multi-*WBAN* concurrent scheduling is then formulated as a game, and its convergence to NE is shown. Meanwhile, NCL [72] is a low-complexity game-based power control approach is shown to reach NE based on best response and to determine the adjustment in transmission power for each transmission in a *WBAN* according to the interference level.

Noncooperative Games

As pointed out earlier, noncooperative games theory approach proved to better suits TPC in *WBAN*s. In [73], Kazemi et al. propose a noncooperative power control game (NPG) based approach, which considers inter-network interference among nearby *WBAN*s. In NPG, the existence and uniqueness of NE have been shown to match the best response solution. Nonetheless, NPG is more efficient than PAU, discussed above, because it assumes an adaptive power price and factors in the power budget. Unlike PAU and NPG, Kazemi et al. [74] proposed a distributed power control game (GRL) employing reinforcement learning (RL). In GRL, each *WBAN* acts as an

agent and learns from experience to appropriately control the transmission power level in a dynamic environment without any further negotiation and cooperation. In addition, RL results in a better trade-off, i.e., between network utilization and the power constraint for each *WBAN*, than PAU or NPG despite its long convergence time. To expedite convergence, the authors proposed a genetic fuzzy (GA) power controller (FPA) approach [75] that do not require any negotiation among *WBAN*s. FPA requires the SINR and the current transmission power as inputs into GA in order to maximize the capacity and minimize the power consumption and the convergence time. Although FPA outperforms PAU, NPG, and GRL, it does not handle dynamic scenarios in which the interference level is unpredictable. On the other hand, NCR [76] considers the problem of joint relay selection and power control in *WBAN*s. To ensure energy-efficient communication, each sensor seeks a strategy to select its next hop and its transmission power independently in order to ensure short end-to-end delay and jitter.

Nonconventional Games

Unlike the game-based solution discussed above, the social nature, which is germane to *WBAN*s, has been considered in the interference mitigation process. SIP [77] pursues a social interaction and prediction power-based game to mitigate the inter-*WBAN*s interference and maximize their energy savings. Each *WBAN* detects the distance to other *WBAN*s located within its transmission range and informs other reachable *WBAN*s, in order for them to optimize its transmission power and avoid interference. Similarly, Dong et al. [78] proposed a noncooperative social-based game theoretic transmit power control scheme (CPC) to maximize the packet delivery ratio among different coexisting *WBAN*s so that the average transmission power is minimized.

Some game theory interference mitigation schemes opt to provide QoS guarantees. PEG [79] pursues a noncooperative power control game to mitigate the inter-*WBAN* interference. In PEG, the utility function is designed so that the QoS requirement can be met with minimal power consumption for each *WBAN*. To obtain an approximation of the NE point, a noncooperative interference segmentation estimate algorithm has been proposed, which guarantees zero information exchange among the coordinator of *WBAN*s. Similarly, Zhou et al. [80] proposed a game theoretical framework for collision avoidance and time slots allocation for better throughput in *WBAN*s (SAG). The *WBAN* coordinator controls access probability during the contention period based on users' priority and allocates suitable time slots with strategies for best payoff based on link states in GTSs. Whilst, BNC [81] employs a Bayesian noncooperative game for power control. By modeling *WBAN*s as players and active links as types of players in the Bayesian model, BNC tries to maximize each player's expected payoff involving both throughput and energy efficiency without any message passing among *WBAN*s. The uniqueness of Bayesian equilibrium for the game has been derived. Overall, the common problem in game theoretic approaches is that each *WBAN* is modeled as a single link and inter-*WBAN* interference is modeled as a game played among certain fixed links. In practice, since multiple sensors are deployed in different areas of the human body in a *WBAN*, there exist multiple wireless links that differ in

channel gains and cross interference to other *WBAN*s. Basically, *WBAN*s could work independently without coordination and synchronization. A *WBAN* could not be reduced to one single wireless link. In addition, game-based power control protocols do not support the mobility of *WBAN*s. Moreover, game-based protocols do not support QoS and are characterized by long delays. Nonetheless, the majority of them are noncooperative game-based protocols that support the dynamic channels of *WBAN*s and do not require message and information exchange.

4.2.3 Summary of Power Control

The existing link-state based power control interference mitigation protocols have been qualitatively discussed and compared. Various protocols have shown that power control is robust to different body mobility scenarios and suit highly populated environment with *WBAN*s. Although link-state based protocols do not require negotiation and message exchange among *WBAN*s, they significantly consume energy and provide poor QoS support. These protocols are not recommended for dynamic environments, e.g., with a high density of *WBAN*s because the link state varies very frequently due to body posture and movement. On the other hand, game-based power control protocols do not support the mobility of *WBAN*s. Nonetheless, the majority of them are noncooperative game-based protocols that support dynamic channel conditions, e.g., varying channel gain, interference power, etc., and do not require message and information exchange, which reduces the energy consumption across coexisting *WBAN*s. However, game-based protocols do not support QoS and are characterized by long delays. Table 11 provides a comparative summary of the different game-based power control interference mitigation proposals discussed in this subsection.

4.3 Multiple Access

With contention-based MAC, e.g., CSMA, sensors within a *WBAN* can decide their medium access pattern without the need for any synchronization. However, the individual performance of these sensors may be degraded because of the high overhead incurred due to addressing the medium access collisions when the density of *WBAN*s is high. Such overhead is mainly due to time and energy consumption during a backoff. On the other hand, contention-free protocols use time synchronization to achieve collision-free transmissions and high throughput. However, one of the major limitations of contention-free approach is the need for time synchronization which is very costly to achieve in the distributed networks, e.g., coexisting *WBAN*s. Another limitation is the time slot allocation, which becomes challenging particularly when the different coexisting *WBAN*s employ non-similar duty cycles, i.e., the number of active sensors in a period of time is not consistent among these *WBAN*s. Contention-free MAC is reliable and energy-efficient [9, 82] in relatively low-density *WBAN*s,

Table 11 Comparison of game-based power control interference mitigation proposals for *WBAN*s

	THR	EC	REL	MOB	COP	CHST	QoS	TOF	CNV
CSR [70]	High	Med	Med	No	Yes	Dyna	No	No	Slow
PAU [71]	Low	High	Low	No	Yes	Static	No	Yes	Slow
NCL [72]	High	Low	High	No	Yes	Dyna	No	No	Fast
NPG [73]	High	Low	High	No	No	Dyna	Yes	Yes	Slow
GRL [74]	High	Low	Med	No	No	Dyna	No	Yes	Slow
FPA [75]	Med	Low	High	No	No	Dyna	No	No	Fast
NCR [76]	Low	Low	Low	Yes	No	Dyna	Yes	Yes	Slow
SIP [77]	High	Low	High	No	Yes	Dyna	Yes	No	Slow
CPC [78]	High	Low	High	No	No	Dyna	No	No	Slow
PEG [79]	High	Low	Med	No	No	Dyna	Yes	No	Fast
SAG [80]	High	Med	High	No	No	Dyna	Yes	No	Fast
BNC [81]	High	High	High	No	No	Dyna	No	Yes	Slow

though extra energy is consumed for their periodic synchronization and control messages.

The *IEEE 802.15.6* MAC protocol does not support all the requirements of *WBAN*s. Within the superframe structure defined in [4], there are some periods of time that are not occupied most of the time, which limits the channel utilization. Sensors that do not have data to transmit should wake-up periodically to receive beacon frames, which increases their energy consumption. Body movements may cause deep fading that lasts for up to 400 ms [14, 83]. However, deep fading is not considered in the *IEEE 802.15.6* MAC design as well as TDMA ordering is kept fixed in the superframe, which both cause packet losses and reduce the reliability of *WBAN*s. Importantly, the *IEEE 802.15.6* standard [4] does not mandate a specific MAC design that considers the heterogeneous traffic, the mobility and the dynamic channel in *WBAN*s. In addition, the standard focuses only on intra-*WBAN* communication. Such flexibility in the *IEEE 802.15.6* standard has motivated quite a few studies for mitigating inter-*WBAN* interference at the MAC level, as we discussed below.

4.3.1 Superframe Modification

Some of the published protocols have pursued superframe modification in order to diminish the probability of medium access collision. These protocols basically modify the internal structures and their ordering as well as the size of the superframe in order to provide energy-efficient and reliable communication for *WBAN*s. ASL [84] is an example of these adaptive MAC protocols which opts to reduce the energy consumption and improve the throughput. ASL employs CSMA/CA to adjust superframe length according to the level of interference. Whereas, in [45], a novel transition matrix method to estimate the channel dynamics has been proposed. Based on channel dynamics estimation, Zhou et al. have revealed the fundamental effect of a proper superframe length in opportunistic scheduling and further designed a simple scheduling scheme, namely, QSM, that dynamically adjusts the superframe length according to the channel condition. While, DIM [85] adjusts the length between superframe's scheduling phase (SP) and contention access phase dynamically according to the different levels of interference. In essence, the length of SP will be reduced when the channel utilization in SP decreases and will be expanded on the contrary. On the other hand, RAP [38] is a MAC protocol based on adaptive resource allocation and traffic prioritization for *WBAN*s. RAP adaptively modifies the interval of the consecutive transmissions according to the medical status of the *WBAN* user and the channel conditions. Moreover, RAP employs a synchronization method to keep sensors sleeping as long as they do not have pending data to transmit in order to save energy. However, one critical challenge for *WBAN*s is to maintain an accepted level of QoS, while at the same time, ensuring energy-efficient communication. CAC [86] is a TDMA-based MAC protocol that addresses such challenge. CAC dynamically adjusts the transmission order and transmission duration of sensors based on mobility-incurred channel status and traffic characteristics of *WBAN*s. In addition, the time slot allocation is further optimized by minimizing energy consumption and synchronization overhead of sensors subject to QoS constraints. *Overall, as the majority of the protocols discussed in this subsection decide the modification of the superframe according to channel conditions, e.g., SINR, the protocols are unsuitable for highly mobile and relatively high density WBANs, as the channel condition changes very quickly, which makes their implementation inefficient.*

4.3.2 Superframe Interleaving

One way to limit the probability of collision is through superframe interleaving. Basically, the coordinators of *WBAN*s exchange information in order to prevent the active periods of their corresponding superframes from overlapping with each other. CST [87] pursues the simplest and most intuitive, yet inefficient solution by creating a common TDMA medium access schedule among multiple coexisting *WBAN*s in order to mitigate the interference and in consequence improve the throughput. CST defines the time when the coordinators exchange their individual medium access schedules. DCD [26] is a distributive approach where the coordinators

collaboratively rearrange the active durations of the superframes of the coexisting *WBAN*s. Accordingly, only those *WBAN*s which can coexist in the channel adjust their superframes. DCD is found to be effective in minimizing interference and improving channel utilization. Meanwhile, FBS [88, 89] is a distributed TDMA-based beacon interval shifting protocol to reduce the packet loss, power consumption, and data delivery latency. FBS prevents the wake-up period of a *WBAN* from overlapping with the wake-up periods of other coexisting *WBAN*s, by employing carrier sensing at the coordinator before a beacon transmission. Grassi et al. [25] used centralized multiple access mechanisms that reschedule beacons to avoid active period overlap, reducing the interferences among distinct *WBAN*s (B^2R). While, AIA [69] employs a distributed asynchronous inter-*WBAN* interference avoidance scheme based on both CSMA/CA and TDMA. AIA includes the timing offset and dynamically adjusts the schedule of the TDMA period to avoid collisions when such period overlaps with those of between nearby *WBAN*s. AIA adapts to the level of interference in multiple mobile *WBAN* environments as well as improves the coordination time without incurring significant complexity overhead. *Overall, such time-sharing based solutions in which WBANs interleave their active period through negotiation or contention are ineffective when the load in WBANs is heavy and duty cycle of WBAN is high. The rescheduling may cause significant transmission delay if there are a large number of coexisting WBANs.*

4.3.3 Hybrid Solutions

Some interference mitigation solutions have pursued a hybrid contention-free and contention-based approach in order to leverage their advantages. 2LM [3] is a two-layer based MAC protocol in which the coordinator of the *WBAN* schedules transmissions within its *WBAN* using TDMA, and employs a carrier sensing mechanism to deal with inter-*WBAN* collisions. 2LM reduces transmission collisions, delay, and energy. However, 2LM is not adaptive to the interference level and does not specify any sleeping mechanism to avoid unnecessary wake-up and the delay due to the long back-off. While, HEH [46] is a hybrid polling MAC protocol leverages harvested energy from the human body. HEH combines polling and probabilistic contention access methods in order to enable prioritized medium access to the sensors. HEH improves the *WBAN* energy efficiency, throughput, and QoS. QoM [90] is a QoS-based MAC designed for heterogeneous high-traffic *WBAN*s. QoM employs preemptive priority scheduling mechanism among *WBAN*s and a fuzzy inference within a *WBAN* to avoid interference. QoM does not consider the dynamic environment, changing topology, and high density of *WBAN*s. Meantime, ISM [91] is an adaptive and energy-efficient multichannel MAC protocol based on channel hopping for *WBAN*s. ISM employs a collision prevention mechanism, a coordinator node rotation mechanism, and a transmission power adjustment method to reduce the delay and energy consumption, as well as improve the throughput. It is worth noting that a star topology is employed by all multiple access based interference

mitigation protocols discussed in this subsection. *Hybrid* denotes a mix of TDMA and CSMA is employed by an interference mitigation protocol, as explained above.

4.3.4 Summary of Medium Access Based Mitigation Protocols

The published work on MAC protocols has demonstrated that due to their flexibility, contention-based protocols cope better with distributed networks, which make them possible solutions for *WBAN* applications. Whilst, contention-free, e.g., TDMA, could be one possible solution to avoid intra-*WBAN* interference [92]. Contention-free and contention-based approaches are suitable for relatively low density of *WBAN*s with low-occupancy channels and few sensors [93, 94]. However, these approaches are unsuitable for highly mobile *WBAN*s with high density of sensors and with high-traffic load as these approaches impose significant medium access collision problems and long delays, due to the channel condition that changes very quickly, and hence their implementation becomes inefficient. Particularly, time-sharing based solutions in which *WBAN*s interleave their active period through negotiation or contention are ineffective when the load in *WBAN*s is heavy and duty cycle of *WBAN* is high. The rescheduling may cause significant transmission delay if there are a large number of coexisting *WBAN*s. Table 12 provides a comparative summary of the different multiple access interference mitigation proposals discussed in this subsection.

4.4 Link Adaptation

Interference, in essence, affects the individual wireless links. Therefore, one way to mitigate the effect of interference is to adjust the link parameters. For example, the link data rate could change based on the SINR of the channel. Link adaptation protocols opt to dynamically match the modulation and coding parameters, namely, data rate, modulation scheme, etc., of the transmitted signal to the channel conditions (e.g., the path loss, the co-channel interference coming from other nearby transmitters, the available transmitter power margin, RSSI, etc). These protocols invariably require some channel state information at the transmitter. In the balance of this section, we provide an overview of published link adaptation schemes in the realm of *WBAN*s.

4.4.1 Data Rate Adjustment

The implementation of TPC mechanism is very challenging in dynamic scenarios when the SINR varies very frequently due to *WBAN*s mobility. Data rate adjustment protocols, on the other hand, are simple to implement and are able to preserve an acceptable link quality in high-level interference environments.

Yang et al. [47] proposed several interference mitigation schemes (MRC) such as adaptive modulation, data rates, and duty cycles to preserve acceptable link quality.

Table 12 Comparison of published multiple access interference mitigation proposals for *WBAN*s

	EC	CHUT	THR	DEL	REL	QoS	COP	CEX	MOB	MAC
2LM [3]	Med	High	Med	High	Med	Med	No	Yes	No	Hybrid
B^2R [25]	Med	Med	Med	Low	Med	Med	No	Yes	No	CSMA
DCD [26]	High	Low	Low	High	Low	Low	Yes	Yes	No	CSMA
AIA [69]	Med	Med	Med	Low	Low	Low	No	Yes	Yes	Hybrid
ASL [84]	Med	High	Low	Low	Med	Med	No	Yes	No	CSMA
OSM [45]	Low	High	High	High	Med	High	No	No	No	TDMA
DIM [85]	High	Med	Low	High	Low	Low	No	Yes	No	CSMA
RAP [38]	Med	Med	Low	Med	Med	Med	No	No	Yes	Hybrid
CAC [86]	Med	Med	Med	Med	Med	Med	No	Yes	Yes	Hybrid
CST [87]	Low	N/A	Med	High	Low	Low	Yes	Yes	NO	TDMA
FBS [88, 89]	Low	High	High	Low	High	High	No	Yes	No	TDMA
HEH [46]	Low	High	High	Low	High	High	No	No	Yes	TDMA
QoM [90]	Med	Med	Med	High	Low	Med	No	Yes	No	Hybrid
isM [91]	High	Med	Med	Med	Med	Med	No	Yes	Yes	CSMA

The *WBAN*'s coordinator selects the appropriate scheme for the sensors based on the level of experienced interference. Similarly, LAC [48] is a contention-based link adaption scheme for interference mitigation within a single *WBAN*. In LAC, the sensors employ link adaptation strategy to select the appropriate modulation scheme like adaptive data rate to decrease the packet error rate according to the experienced channel quality and level of interference. On the other hand, Moungla et al. [95] proposed a multi-hop tree-based *WBAN* topology design that assumes the mobility of the *WBAN* while ensuring reliable data delivery. In such a design, which is called TDM, the *WBAN* sensors share a small number of channels, whereas, the relay nodes share the most number of channels in order to improve the data flow across the *WBAN*. To mitigate co-channel interference among relays, adaptive data rate and duty cycle are used.

Table 13 Comparison of published data rate adjustment interference mitigation proposals for *WBAN*s

	DR	EC	THR	PER	CHP	TPO	MAC	MOB	CEX
MRC [47]	High	Low	High	Low	SINR	Star	TDMA	No	Yes
LAC [48]	Med	Med	Med	Low	SNR	Star	CSMA	No	No
TDM [95]	High	Low	High	Low	SINR	2-hop	TDMA	Yes	No

Overall, data rate adjustment protocols are effective and simple to implement and can achieve acceptable link quality. However, they do not suite highly mobile and densely deployed WBANs because of the fast-changing channel conditions, e.g., SINR. Table 13 provides a comparative summary of the different data rate adjustment interference mitigation proposals discussed in this subsection.

4.4.2 Two-Hop Communication

Due to the severe signal attenuation and shadowing effect caused by the human body itself and its mobility effects, deploying one-hop star topology could be ineffective in achieving reliable communication in *WBAN*s. As stated in the *IEEE 802.15.6* standard [4], the two-hop communication is a very promising solution as it exploits the spatial diversity to improve *WBAN* energy efficiency and reliability. Moreover, using the two-hop communication allows for better *WBAN* interference mitigation and coexistence by increasing the SINR of data packets received at the coordinator. Despite these advantages, two-hop communications may shorten the energy lifetime of the relay nodes. The frequent relaying process through the same set of relay nodes will quickly drain their batteries, which will make them die faster than other sensor nodes in the network. In [96], the performance of one-hop and two-hop communications schemes is compared to demonstrate the effectiveness of the relay transmission mechanism in *WBAN*s. Figure 9 shows an example for one-hop and two-hop communication schemes [97].

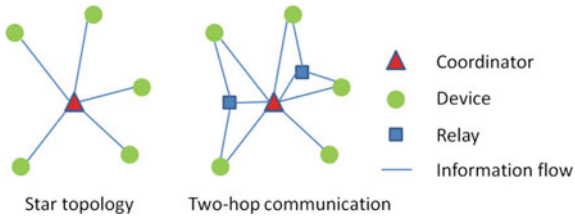


Fig. 9 One-hop and two-hop communication schemes

Several interference mitigation solutions based on two-hop communication have been published for *WBAN*s. Feng et al. [39] presented temporal and spatial correlation models to better characterize the slow fading effect of on-body channels. They proposed a dynamic prediction-based relay transmission scheme (PRT) that makes full use of the correlation characteristics of on-body channels and achieves improvement in the energy efficiency and transmission reliability in a *WBAN*. In PRT, when to relay and which node to become a relay are decided in an optimal way based on the last known channel states. PRT needs neither extra signaling procedure nor dedicated channel sensing period. DMT [98] is a decode with merge technique (DMT) that maintains the relaying mode by merging frames from relayed and generated by relay nodes in order to increase the throughput at the *WBAN* coordinator without increasing the energy consumption. However, DMT does not address the interference occurring at relay nodes. While, Dong et al. [97] proposed a relay-assisted cooperative communications scheme (LRS) for a *WBAN*. LRS considers two relay nodes and provides a 3-link diversity gain (DG) to the coordinator with selection combining (SC). Similarly, SOR [40] is a two-hop scheme integrated with opportunistic relaying (OR) for mobile *WBAN*s (SOR). By using received SINR at the relay and coordinator nodes, SOR chooses the best relay to decode and forward the data at the same time with the direct link. JNT [99] is a two-hop cooperative scheme integrated with transmit power control (JNT) and based on simple channel prediction for *WBAN*s. In JNT, a transmit power control mechanism is integrated into sensor and relay nodes to prolong sensor battery lifetime and mitigate the interference at the *WBAN* coordinator. Meantime, DFP [23] is a TDMA-based two-hop communication scheme (DFP) among multiple mobile non-coordinated *WBAN*s. The coordinator of *WBAN*-of-interest used a decode-and-forward protocol with two dual-hop links, two relays and selection combining. Then, a suitable TDMA scheme is used to allocate time slots for each link packet transmission.

Overall, using two-hop interference mitigation based protocols improve the channel gain and SINR thresholds at low outage probability which further increases the throughput at WBAN receivers. Moreover, these protocols reduce the level crossing rate (LCR) at low SINR values, e.g., an LCR of 1 Hz, an SINR threshold value increases by 6 dB, and extend the average non-fade duration that both lower the overhead for scheduling transmissions [97]. *Level crossing rate (LCR) is a statistic that describes the measure of the rapidity of the fading and quantifies how often the fading crosses some threshold. While, the non-fade duration quantifies how long the signal spends above some threshold, where there exists sufficient signal strength during which the receiver can work reliably and at low bit error rate. Therefore, using two-hop based protocols allows for more packets of large size to be transmitted, i.e., larger data rates, which reduce the transmission delay. However, two-hop transmission may also introduce some additional latency to the packet delivery that may be unacceptable in time-sensitive healthcare applications, e.g., heart vital data.*

Table 14 Comparison of published two-hop based interference mitigation proposals for *WBAN*s

	EC	REL	OP	LCR	THR	CHP	MAC	MOB	CEX
DFP [23]	N/A	High	High	N/A	High	SINR	TDMA	Yes	Yes
PRT [39]	Low	High	High	Low	High	SINR	TDMA	Yes	No
DMT [98]	Low	Low	N/A	N/A	High	LQI	CSMA	No	No
LRS [97]	N/A	High	High	Low	High	Gain	CSMA	Yes	Yes
SOR [40]	N/A	High	High	Low	High	SINR	TDMA	Yes	Yes
JNT [99]	Low	High	High	N/A	High	SINR	TDMA	No	Yes

4.4.3 Summary of Link Adaptation

Link adaptation protocols do not suit highly mobile and densely deployed *WBAN*s because of the fast-changing channel conditions, e.g., SINR. On the other hand, as these protocols may exploit the diversity gain, the co-channel interference could be better mitigated, by increasing the SINR and the channel gain at receivers, which further increases the throughput. However, these protocols may introduce some additional energy cost at relays and latency to the packet delivery that may be unacceptable in time-sensitive healthcare applications. Table 14 provides a comparative summary of the different two-hop based interference mitigation proposals discussed in this subsection.

In summary, based on our study to qualitatively compare interference mitigation techniques for *WBAN*s, it can be concluded that there is no dominating technique that outperforms the others. In dynamic channels, there is a need for a trade-off between network throughput and power consumption. In addition, the network lifetime, reliability, delay (latency), QoS, mobility, and channel dynamics must also be considered with the design. Moreover, the existing interference mitigation techniques do not completely address QoS requirements and achieve the desired performance in some healthcare applications. We envision that cross-layer based interference mitigation protocols will be a promising solution methodology that is worthy increased attention.

5 Conclusions and Future Research Directions

In this chapter, we have presented a brief overview of communication architecture and various technological protocols designed for *WBAN* systems. The issues related to the coexistence among *WBAN*s and between *WBAN*s and other wireless networks have

been analyzed. A comparative review of the radio co-channel interference mitigation and avoidance techniques in the literature has been provided. These techniques are categorized as resource allocation, power control, and some solutions which are based on the incorporation of multiple medium access arbitration mechanisms and link adaptation. The purpose of these techniques is to ensure that *WBAN*s operate in a stable way even in the populated area, under high mobility conditions, and high levels of co-channel interference. We summarize the advantages and disadvantages of each technique as follows:

- Resource allocation protocols, namely, channels and time, may work efficiently in dynamic environments, i.e., highly mobile and densely deployed *WBAN*s, and under high interference conditions. In general, with the exception of graph-based schemes, these protocols may support an acceptable level of QoS requirements.
- Our study demonstrates that link-state based power control protocols do not require negotiation and message exchange among *WBAN*s. However, these protocols are not recommended for dynamic environments, e.g., with a high density of *WBAN*s, because the link state varies very frequently due to body posture and movement. On the other hand, game-based power control protocols do not support the mobility of *WBAN*s. Nonetheless, the majority of them are noncooperative game-based protocols that support dynamic channel conditions and do not require message and information exchange, which reduce the energy consumption across coexisting *WBAN*s. However, game-based protocols do not support QoS and are characterized by long delays.
- The published work on MAC protocols have demonstrated that either and contention-free or contention-based approaches are suitable for relatively low density of *WBAN*s with low-occupancy channels and few sensors. However, neither of them could efficiently handle highly mobile *WBAN*s with the high-traffic load. A hybrid approach that combines the advantages of both methodologies is deemed promising.
- Data rate adjustment protocols do not suite highly mobile and densely deployed *WBAN*s because of the fast-changing channel conditions, e.g., SINR. In addition, these protocols may introduce some additional energy cost at relays and latency to the packet delivery that may be unacceptable in time-sensitive healthcare applications. On the other hand, using two-hop interference mitigation based protocols improves the channel gain and SINR which further increases the throughput at *WBAN* receivers. Using two-hop based protocols allows for more packets of large size to be transmitted, i.e., larger data rates, which reduce the transmission delay. However, the two-hop transmission may also introduce some additional latency to the packet delivery that may be unacceptable in time-sensitive healthcare applications, e.g., heart vital data.

Although our study qualitatively compares these techniques for *WBAN*s and provides important insights about them, we arrive at the conclusion that there is no dominating technique that outperforms the others. Moreover, the existing interference mitigation techniques do not completely address QoS requirements and achieve the desired performance in some healthcare applications. Furthermore, we show that

existing solutions fall short from achieving satisfactory performance, and warrant more investigation by the research community. We envision that cross-layer based interference mitigation protocols will be a promising solution methodology that is worth increased attention.

References

1. Latré, B., Braem, B., Moerman, I., Blondia, C., Demeester, P.: A survey on wireless body area networks. Wirel. Netw. **17**(1), 1–18 (2011)
2. Samaneh, M., Mehran, A., Justin, L., David, S., Abbas, J.: Wireless body area networks: A survey. IEEE Commun. Surv. Tutor. **16**(3), 1658–1686 (2014)
3. Chen, G.T., Chen, W.T., Shen, S.H.: 2l-mac: A mac Protocol with Two-Layer Interference Mitigation in Wireless Body Area networks for Medical Applications, pp. 3523–3528 (2014)
4. IEEE standard for local and metropolitan area networks—part 15.6: Wireless body area networks. IEEE Std 802.15.6-2012
5. Ullah, S., Khan, P., Ullah, N., Saleem, S., Higgins, H., Kwak, K.S.: A review of wireless body area networks for medical applications. IJCNS **2**(8), 797–803 (2010)
6. Smith, D.B., Miniutti, D., Lamahewa, T.A., Hanlen, L.W.: Propagation models for body-area networks: A survey and new outlook. IEEE Antennas Propag. Mag. **55**(5), 97–117 (2013)
7. Sukor, M., Ariffin, S., Fisal, N., Yusof, S.K.S., Abdallah, A.: Performance study of wireless body area network in medical environment. In: 2008 Second Asia International Conference on Modelling x00026; Simulation (AMS), pp. 202–206 (2008)
8. Natarajan, A., Motani, M., de Silva, B., Yap, K.-K., Chua, K.C.: Investigating network architectures for body sensor networks. In: Proceedings of the 1st ACM SIGMOBILE International Workshop on Systems and Networking Support for Healthcare and Assisted Living Environments, HealthNet '07, pp. 19–24. ACM, New York, NY, USA (2007)
9. Movassaghi, S., Majidi, A., Jamalipour, A., Smith, D., Abolhasan, M.: Enabling interference-aware and energy-efficient coexistence of multiple wireless body area networks with unknown dynamics. IEEE Access **4**, 2935–2951 (2016)
10. IEEE standard for information technology—local and metropolitan area networks—specific requirements– part 15.1a: Wireless medium access control (mac) and physical layer (phy) specifications for wireless personal area networks (wpan). IEEE Std 802.15.1-2005 (Revision of IEEE Std 802.15.1-2002), pp. 1–700 (2005)
11. Bluetooth low energy technology. http://www.bluetooth.com/Pages/low-energy-tech-info.aspx
12. IEEE standard for local and metropolitan area networks–part 15.4: Low-rate wireless personal area networks (lr-wpans). IEEE Std 802.15.4-2011 (Revision of IEEE Std 802.15.4-2006)
13. IEEE standard for local and metropolitan area networks—part 15.6: Wireless body area networks
14. Yazdandoost, K.Y., Sayrafian-Pour, K.: Channel model for body area network (ban). Networks, p. 91 (2009)
15. Yuce, M.R., Ng, S.W.P., Myo, N.L., Lee, C.K., Khan, J.Y., Liu, W.: A mics band wireless body sensor network. In: 2007 IEEE Wireless Communications and Networking Conference, pp. 2473–2478 (2007)
16. Khan, I., Nechayev, Y.I., Hall, P.S.: On-body diversity channel characterization. IEEE Trans. Antennas Propag. **58**(2), 573–580 (2010)
17. Cotton, S.L., Scanlon, W.G., Guy, J.: The $\kappa - \mu$ distribution applied to the analysis of fading in body to body communication channels for fire and rescue personnel. IEEE Antennas Wirel. Propag. Lett. **7**, 66–69 (2008)

18. Smith, D., Hanlen, L., Zhang, J., Miniutti, D., Rodda, D., Gilbert, B.: Characterization of the dynamic narrowband on-body to off-body area channel. In: 2009 IEEE International Conference on Communications, pp. 1–6, (2009)
19. Scanlon, W.G., Cotton, S.L.: Understanding on-body fading channels at 2.45 ghz using measurements based on user state and environment. In: 2008 Loughborough Antennas and Propagation Conference, pp. 10–13 (2008)
20. Fort, A., Desset, C., Wambacq, P., Biesen, L.V.: Indoor body-area channel model for narrowband communications. IET Microw. Antennas Propag. **1**(6), 1197–1203 (2007)
21. Smith, D., Hanlen, L., Miniutti, D., Zhang, J., Rodda, D., Gilbert, B.: Statistical characterization of the dynamic narrowband body area channel. In: 2008 First International Symposium on Applied Sciences on Biomedical and Communication Technologies, p. 1–5, Oct 2008
22. Guraliuc, A.R., Serra, A.A., Nepa, P., Manara, G.: Channel model for on body communication along and around the human torso at 2.4 GHz and 5.8 GHz. In: 2010 International Workshop on Antenna Technology (iWAT), pp. 1–4 (2010)
23. Dong, J., Smith, D.: Cooperative body-area-communications: enhancing coexistence without coordination between networks. In: 2012 IEEE 23rd International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp. 2269–2274, Sept 2012
24. Abi research: Wearable computing devices, like apple's iwatch, will exceed 485 million annual shipments by 2018. https://www.abiresearch.com/press/wearable-computing-devices-like-apples-iwatch-will. Accessed Dec 2014
25. Grassi, P.R., Rana, V., Beretta, I., Sciuto, D.: b^2irs: A technique to reduce ban-ban interferences in wireless sensor networks. In: 2012 Ninth International Conference on Wearable and Implantable Body Sensor Networks, pp. 46–51 (2012)
26. Deylami, M., Jovanov, E.: A distributed and collaborative scheme for mitigating coexistence in ieee 802.15.4 based wbans. In: Proceedings of the 50th Annual Southeast Regional Conference, ACM-SE '12, pp. 1–6. ACM, New York, NY, USA (2012)
27. Wang, X., Cai, L.: Interference analysis of co-existing wireless body area networks. In: 2011 IEEE Global Telecommunications Conference (GLOBECOM 2011), pp. 1–5, Dec 2011
28. Tan, L., Feng, Z., Li, W., Jing, Z., Gulliver, T.A.: Graph coloring based spectrum allocation for femtocell downlink interference mitigation. In: 2011 IEEE Wireless Communications and Networking Conference, pp. 1248–1252 (2011)
29. Chang, R.Y., Tao, Z., Zhang, J., Kuo, C.C.: A graph approach to dynamic fractional frequency reuse (ffr) in multi-cell ofdma networks. In: 2009 IEEE International Conference on Communications, pp. 1–6 (2009)
30. Pateromichelakis, E., Shariat, M., Quddus, A.U., Tafazolli, R.: On the evolution of multi-cell scheduling in 3g pplte/lte-a. IEEE Commun. Surv. Tutor. **15**(2), 701–717 (2013)
31. Fan, P., Min, G., Pan, Y. (eds.): Design fundamentals and interference mitigation for cellular networks. In: Advances in Wireless Networks: Performance Modelling, Analysis and Enhancement (2008)
32. Fantacci, R.: Proposal of an interference cancellation receiver with low complexity for ds/cdma mobile communication systems. IEEE Trans. Veh. Technol. **48**(4), 1039–1046 (1999)
33. Barakah, D.M., Ammad-uddin, M.: A survey of challenges and applications of wireless body area network (wban) and role of a virtual doctor server in existing architecture. In: 2012 Third International Conference on Intelligent Systems, Modelling and Simulation (ISMS), pp. 214–219, Feb 2012
34. Javaid, N., Khan, N.A., Shakir, M., Khan, M.A., Hussain Bouk, S., Khan, Z.A.: Ubiquitous healthcare in wireless body area networks-a survey (2013). arXiv:1303.2062
35. Movassaghi, S., Abolhasan, M., Smith, D.: Interference mitigation in wbans: challenges and existing solutions. In: Workshop on Advances in Real-time Information Networks (2013)
36. Yan, Z., Liu, B., Chen, C.W.: Qos-driven scheduling approach using optimal slot allocation for wireless body area networks. In: 2012 IEEE 14th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 267–272 (2012)
37. Quwaider, M., Rao, J., Biswas, S.: Body-posture-based dynamic link power control in wearable sensor networks. IEEE Commun. Mag. **48**(7), 134–142 (2010)

38. Rezvani, S., Ghorashi, S.A.: Context aware and channel-based resource allocation for wireless body area networks. IET Wirel. Sens. Syst. **3**(1), 16–25 (2013)
39. Feng, H., Liu, B., Yan, Z., Zhang, C., Chen, C.W.: Prediction-based dynamic relay transmission scheme for wireless body area networks. In: 2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp. 2539–2544, Sept 2013
40. Dong, J., Smith, D.: Opportunistic relaying in wireless body area networks: Coexistence performance. In: 2013 IEEE International Conference on Communications (ICC), pp. 5613–5618. IEEE (2013)
41. Liang, S., Ge, Y., Jiang, S., Tan, H.P.: A lightweight and robust interference mitigation scheme for wireless body sensor networks in realistic environments. In: 2014 IEEE Wireless Communications and Networking Conference (WCNC), pp. 1697–1702 (2014)
42. Movassaghi, S., Abolhasan, M., Smith, D., Jamalipour, A.: Aim: adaptive internetwork interference mitigation amongst co-existing wireless body area networks. In: 2014 IEEE Global Communications Conference (GLOBECOM), pp. 2460–2465 (2014)
43. See, T.S.P., Ge, Y., Chiam, T.M., Kwan, J.W., Kim, C.W.: Experimental correlation of path loss with system performance in wban for healthcare applications. In: 2011 IEEE 13th International Conference on e-Health Networking, Applications and Services, pp. 221–224 (2011)
44. Ge, Y., Kwan, J.W., Pathmasuntharam, J.S., Di, Z., See, T.S.P., Ni, W., Kim, C.W., Chiam, T.M., Ma, M.: Performance benchmarking for wireless body area networks at 2.4 GHz. In: 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 2249–2253, Sept 2011
45. Zhou, Y., Sheng, Z., Leung, V.C.M., Servati, P.: Beacon-based opportunistic scheduling in wireless body area network. In: 2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), pp. 4995–4998, Aug 2016
46. Ibarra, E., Antonopoulos, A., Kartsakli, E., Verikoukis, C.: Heh-bmac: Hybrid polling mac protocol for wbans operated by human energy harvesting. Telecommun. Syst. **58**(2), 111–124 (2015)
47. Yang, W.-B., Sayrafian-Pour, K.: Interference mitigation using adaptive schemes in body area networks. Int. J. Wirel. Inf. Netw. **19**(3), 193–200 (2012)
48. Martelli, F., Verdone, R., Buratti, C.: Link adaptation in ieee 802.15.4-based wireless body area networks. In: 2010 IEEE 21st International Symposium on Personal, Indoor and Mobile Radio Communications Workshops (PIMRC Workshops), pp. 117–121, Sept 2010
49. Thepvilojanapong, N., Motegi, S., Idoue, A., Horiuchi, H.: Adaptive channel and time allocation for body area networks. IET Commun. **5**(12), 1637–1649 (2011)
50. Movassaghi, S., Abolhasan, M., Smith, D.: Smart spectrum allocation for interference mitigation in wireless body area networks. In: 2014 IEEE International Conference on Communications (ICC), pp. 5688–5693. IEEE (2014)
51. Ali, M.J., Moungla, H., Younis, M., Mehaoua, A.: Efficient medium access arbitration among interfering WBANs using latin rectangles. Ad Hoc Netw. (2018)
52. Donald Keedwell, A., Dénes, J.: Latin Squares and Their Applications. North-Holland, Boston, MA, USA (2015)
53. Cheng, S.H., Huang, C.Y.: Coloring-based inter-wban scheduling for mobile wireless body area networks. IEEE Trans. Parallel Distrib. Syst. **24**(2), 250–259 (2013)
54. Movassaghi, S., Abolhasan, M., Smith, D.: Cooperative scheduling with graph coloring for interference mitigation in wireless body area networks. In: IEEE International Conference on WCNC (2014)
55. Xie, Z., Huang, G., He, J., Zhang, Y.: A clique-based wban scheduling for mobile wireless body area networks. Proc. Comput. Sci. **31**, 1092–1101 (2014)
56. Seo, S., Bang, H., Lee, H.: Coloring-based scheduling for interactive game application with wireless body area networks. J. Supercomput. **72**(1), 185–195 (2016)
57. Huang, W., Quek, T.Q.S.: On constructing interference free schedule for coexisting wireless body area networks using distributed coloring algorithm. In: IEEE International Conference on BSN (2015)

58. Movassaghi, S., Majidi, A., Smith, D., Abolhasan, M., Jamalipour, A.: Exploiting unknown dynamics in communications amongst coexisting wireless body area networks. In: 2015 IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2015)
59. Pantazis, N.A., Vergados, D.D.: A survey on power control issues in wireless sensor networks. Commun. Surv. Tutor. **9**(4), 86–107 (2007)
60. Xiao, M., Shroff, N.B., Edwin, K.P.: Chong. A utility-based power-control scheme in wireless cellular systems. IEEE/ACM Trans. Netw. **11**(2), 210–221 (2003)
61. Cagalj, M., Ganeriwal, S., Aad, I., Hubaux, J.P.: On selfish behavior in csma/ca networks. In: Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies, vol. 4, pp. 2513–2524 (2005)
62. Son, D., Krishnamachari, B., Heidemann, J.: Experimental study of the effects of transmission power control and blacklisting in wireless sensor networks. In: 2004 First Annual IEEE Communications Society Conference on Sensor and Ad Hoc Communications and Networks, IEEE SECON 2004, pp. 289–298 (2004)
63. Lin, S., Zhang, J., Zhou, G., Gu, L., Stankovic, J.A., He, T.: Atpc: Adaptive transmission power control for wireless sensor networks. In: Proceedings of the 4th International Conference on Embedded Networked Sensor Systems, SenSys '06, pp. 223–236. ACM, New York, NY, USA (2006)
64. Guan, T., Yi, C., Qiao, D., Xu, L., Li, Y.: Pid-based transmission power control for wireless body area network. In: 2014 12th International Conference on Signal Processing (ICSP), pp. 1643–1648 (2014)
65. Xiao, S., Dhamdhere, A., Sivaraman, V., Burdett, A.: Transmission power control in body area sensor networks for healthcare monitoring. IEEE J. Sel. Areas Commun. **27**(1), 37–48 (2009)
66. Kim, S., Eom, D.S.: Link-state-estimation-based transmission power control in wireless body area networks. IEEE J. Biomed. Health Inform. **18**(4), 1294–1302 (2014)
67. Movassaghi, S., Abolhasan, M., Smith, D., Jamalipour, A.: Joint energy harvesting and internetwork interference mitigation amongst coexisting wireless body area networks. In: ICST, 11 (2014)
68. Chen, Q., Su, C., Zhang, H., Chai, R.: User service oriented power allocation algorithm for wireless body area sensor networks. In: 5th IET International Conference on Wireless, Mobile and Multimedia Networks (ICWMMN 2013), pp. 37–40 (2013)
69. Kim, E.-J., Youm, S., Shon, T., Kang, C.-H.: Asynchronous inter-network interference avoidance for wireless body area networks. J. Supercomput. **65**(2), 562–579 (2013)
70. Wang, L., Goursaud, C., Nikaein, N., Cottatellucci, L., Gorce, J.M.: Cooperative scheduling for coexisting body area networks. IEEE Trans. Wirel. Commun. **12**(1), 123–133 (2013)
71. Fang, G., Dutkiewicz, E., Yu, K., Vesilo, R., Yu, Y.: Distributed inter-network interference coordination for wireless body area networks. In: 2010 IEEE Global Telecommunications Conference GLOBECOM 2010, pp. 1–5 (2010)
72. Du, D., Hu, F., Wang, F., Wang, Z., Du, Y., Wang, L.: A game theoretic approach for internetwork interference mitigation in wireless body area networks. China Commun. **12**(9), 150–161 (2015)
73. Kazemi, R., Vesilo, R., Dutkiewicz, E., Fang, G.: Inter-network interference mitigation in wireless body area networks using power control games. In: 2010 10th International Symposium on Communications and Information Technologies, pp. 81–86 (2010)
74. Kazemi, R., Vesilo, R., Dutkiewicz, E., Liu, R.P.: Reinforcement learning in power control games for internetwork interference mitigation in wireless body area networks. In: 2012 International Symposium on Communications and Information Technologies (ISCIT), pp. 256–262 (2012)
75. Kazemi, R., Vesilo, R., Dutkiewicz, E.: A novel genetic-fuzzy power controller with feedback for interference mitigation in wireless body area networks. In: 2011 IEEE 73rd Vehicular Technology Conference (VTC Spring), pp. 1–5 (2011)
76. Moosavi, H., Bui, F.M.: Optimal relay selection and power control with quality-of-service provisioning in wireless body area networks. IEEE Trans. Wirel. Commun. **15**(8), 5497–5510 (2016)

77. Zhang, Z., Wang, H., Wang, C., Fang, H.: Interference mitigation for cyber-physical wireless body area network system using social networks. IEEE Trans. Emerg. Top. Comput. **1**(1), 121–132 (2013)
78. Dong, J., Smith, D.B., Hanlen, L.W.: Socially optimal coexistence of wireless body area networks enabled by a non-cooperative game. ACM Trans. Sens. Netw. **12**(4), 26:1–26:18 (2016)
79. Zhao, X., Liu, B., Chen, C., Chen, C.W.: Qos-driven power control for inter-wban interference mitigation. In: 2015 IEEE Global Communications Conference (GLOBECOM), pp. 1–6 (2015)
80. Zhou, J., Guo, A., Xu, J., Nguyen, H., Su, S.: A game theory control scheme in medium access for wireless body area network. In: 10th International Conference on Wireless Communications, Networking and Mobile Computing (WiCOM 2014), pp. 404–409, Sept 2014
81. Zou, L., Liu, B., Chen, C., Chen, C.W.: Bayesian game based power control scheme for inter-wban interference mitigation. In: 2014 IEEE Global Communications Conference (GLOBECOM), pp. 240–245 (2014)
82. Deylami, M.N., Jovanov, E.: A distributed scheme to manage the dynamic coexistence of ieee 802.15.4-based health-monitoring wbans. IEEE J. Biomed. Health. Inform. **18**(1), 327–334 (2014)
83. Hall, P.S., Hao, Y., Nechayev, Y.I., Alomainy, A., Constantinou, C.C., Parini, C., Kamarudin, M.R., Salim, T.Z., Hee, D.T.M., Dubrovka, R., Owadally, A.S., Song, W., Serra, A., Nepa, P., Gallo, M., Bozzetti, M.: Antennas and propagation for on-body communication systems. IEEE Antennas Propag. Mag. **49**(3), 41–58 (2007)
84. Huang, W., Quek, T.Q.S.: Adaptive csma/ca mac protocol to reduce inter-wban interference for wireless body area networks. In: 2015 IEEE 12th International Conference on Wearable and Implantable Body Sensor Networks (BSN), pp. 1–6 (2015)
85. Yuan, B., Liu, J., Liu, W., Zheng, S.: Dim: a novel decentralized interference mitigation scheme in wban. In: 2015 International Conference on Wireless Communications Signal Processing (WCSP), pp. 1–5 (2015)
86. Liu, B., Yan, Z., Chen, C.W.: Medium access control for wireless body area networks with qos provisioning and energy efficient design. IEEE Trans. Mob. Comput. **PP**(99), 1–1 (2016)
87. Mahapatro, J., Misra, S., Manjunatha, M., Islam, N.: Interference mitigation between wban equipped patients. In: 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN), pp. 1–5, Sept 2012
88. Kim, S., Kim, J., Eom, D.: A beacon interval shifting scheme for interference mitigation in body area networks. Sensors **12**(8), 10930–10946 (2012)
89. Kim, S., Kim, S., Kim, J.W., Eom, D.S.: Flexible beacon scheduling scheme for interference mitigation in body sensor networks. In: 2012 9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks (SECON), pp. 157–164 (2012)
90. Jamthe, A., Mishra, A., Agrawal, D.P.: Scheduling schemes for interference suppression in healthcare sensor networks. In: 2014 IEEE International Conference on Communications (ICC), pp. 391–396. IEEE (2014)
91. Kirbas, I., Karahan, A., Sevin, A., Bayilmis, C.: ismac: An adaptive and energy-efficient mac protocol based on multi-channel communication for wireless body area networks. TIIS **7**(8), 1805–1824 (2013)
92. Yang, W.B., Sayrafian-Pour, K.: Interference mitigation for body area networks. In: 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications, pp. 2193–2197, Sept 2011
93. Omeni, O., Wong, A.C.W., Burdett, A.J., Toumazou, C.: Energy efficient medium access protocol for wireless medical body area sensor networks. IEEE Trans. Biomed. Circuits Syst. **2**(4), 251–259 (2008)
94. Tseng, H.W., Sheu, S.T., Shih, Y.Y.: Rotational listening strategy for IEEE 802.15.4 wireless body networks. IEEE Sens. J **11**(9), 1841–1855 (2011)
95. Moungla, H., Jarray, A., Karmouch, A., Mehaoua, A.: Cost-effective reliability-and energy-based intra-wban interference mitigation. In: 2014 IEEE Global Communications Conference, pp. 2399–2404 (2014)

96. Ferrand, P., Maman, M., Goursaud, C., Gorce, J.-M., Ouvry, L.: Performance evaluation of direct and cooperative transmissions in body area networks. Annals of Telecommunications - annales des télécommunications **66**(3), 213–228 (2011)
97. Dong, J., Ge, Y., Smith, D.B.: Two-hop relay-assisted cooperative communication in wireless body area networks: an empirical study. ACM Trans. Sen. Netw. **12**(4):32:1–32:13 (2016)
98. Manirabona, A., Fourati, L.C., Boudjit, S.: Decode and merge cooperative mac protocol for intra wban communication. In: 2014 IEEE 16th International Conference on e-Health Networking, Applications and Services (Healthcom), pp. 146–151 (2014)
99. Dong, J., Smith, D.: Joint relay selection and transmit power control for wireless body area networks coexistence. In: 2014 IEEE International Conference on Communications (ICC), pp. 5676–5681 (2014)

Radiation Control Algorithms in Wireless Networks

Sotiris Nikoletseas, Theofanis P. Raptis, Christoforos Raptopoulos and José Rolim

Abstract Electromagnetic radiation, defined as the total amount of electromagnetic quantity that a target elementary surface is exposed to, is one of the main byproducts from the advancement and wide deployment of wireless distributed systems and ad hoc networks consisting of increasingly more powerful devices and diverse technology. Nevertheless, the extreme benefits of the latter have resulted in the emergence of a new research area in algorithmic network design, with main objective the control of the emitted radiation within such systems. In this chapter, we explore this new research area by presenting two quite distinct approaches for radiation control in wireless distributed systems. In particular, we first study the minimum radiation path problem of finding the lowest radiation trajectory of a person moving from a source to a destination point within the area of a network of wireless devices; this is particularly relevant in smart buildings. Second, we study the problem of efficiently charging a set of rechargeable nodes using a set of wireless energy chargers, under safety constraints on the electromagnetic radiation incurred. For both these problems, we provide hardness indications and theoretical results highlighting interesting structural and algorithmic properties. Furthermore, we present and analyze efficient algorithms and heuristics for approximating optimal solutions, namely minimum

S. Nikoletseas (✉) · C. Raptopoulos
Department of Computer Engineering and Informatics, University of Patras, Computer Technology Institute and Press "Diophantus" (CTI), Patras, Greece
e-mail: nikole@cti.gr

C. Raptopoulos
e-mail: raptopox@ceid.upatras.gr

T. P. Raptis
Institute of Informatics and Telematics, National Research Council, Pisa, Italy
e-mail: theofanis.raptis@iit.cnr.it

J. Rolim
Centre Universitaire d' Informatique, University of Geneva, Geneva, Switzerland
e-mail: jose.rolim@unige.ch

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_20

radiation trajectories and charging schemes, respectively. Finally, we present experimental evidence that not only verifies our theoretical results but also provide new insights that we could not obtain through analysis due to the inherent complexity of these problems.

1 Introduction

Recently, we have witnessed a rapid advance and wide deployment of *wireless distributed systems* (WDS) and *ad hoc networks*, consisting of increasingly more powerful devices and diverse technology such as mobile phones, WiFi, Bluetooth, wireless charging devices, and smart wireless sensory devices. The beneficial use of such technology, however, comes at a price with regards to real-life applications: as a result of the combined operation of such devices, every point of space within the deployment area of such systems is exposed to *electromagnetic radiation* (EMR), i.e., (loosely speaking) electromagnetic quantity produced by wireless devices. Exposure to high electromagnetic radiation has been widely recognized as a *threat to human health*. Its potential risks include, but is not limited to, mental diseases [37], tissue impairment [38], and brain tumor [18]. In addition, there has been solid evidence that pregnant women and children are even more vulnerable to high electromagnetic radiation exposure [10, 14]. We note that, particularly, the radiation levels created by wireless power can be quite high, due to the strength of the electromagnetic fields created. Even if the impact of electromagnetic radiation can be considered controversial, we believe it is worth understanding and control, without however compromising the quality of service offered to the user of wireless communications. For such systems, the broader aim would be to come up with radiation awareness in an adaptive manner, by providing design principles and studying key algorithmic and networking aspects of radiation aware wireless networking. The Computer Science research community has already demonstrated relevant interest from an ICT perspective by considering restrictions in the amount of emitted EMR. This creates a new research area in algorithmic network design for WDS.

This chapter is based on our research papers [35, 36], which concern radiation control in two quite distinct wireless network scenarios.

A. **Minimum radiation paths**. We focus on the minimum radiation path problem of finding the lowest radiation trajectory of a person moving from a source to a destination point within the area of a network of wireless devices. This is particularly relevant in smart buildings; we are aware of the fact that the radiation of tiny wireless sensors is not very high but we view this network type as an example of a broader, heterogeneous wireless network setting. Still, we note that the radiation impact of tiny sensors may be high in wireless body area networks (WBANs) where sensors lie close to vital organs and wearable (or even implanted) sensitive nanoscale medical electronic devices.

In this setting, we first evaluate radiation in well-known topologies (random and grids); randomness is meant to capture not only random placement but also uncertainty of the wireless propagation model. Second, we present several heuristics which provide low radiation paths while keeping path length low; one heuristic gets, in fact, quite close to the offline optimum solution given by a linear program. Finally, we investigate the impact on the heuristics' performance for diverse node mobility.

B. **Low radiation wireless energy transfer**. We study the problem of efficiently charging a set of rechargeable nodes using a set of wireless energy chargers, under safety constraints on the electromagnetic radiation incurred. The particular charging model greatly differs from existing models in that it takes into account real technology restrictions of the chargers and nodes of the system, mainly regarding energy limitations. Our model introduces nonlinear constraints (in the time domain) that radically change the nature of the computational problems we consider.
In this setting, we present and study the *Low Radiation Efficient Charging Problem* (*LREC*), in which we wish to optimize the amount of "useful" energy transferred from chargers to nodes (under constraints on the maximum level of imposed radiation). We present several fundamental properties of this problem and provide indications of its hardness. We propose an iterative local improvement heuristic for LREC, which runs in polynomial time and we evaluate its performance via simulation. Our algorithm decouples the computation of the objective function from the computation of the maximum radiation and also does not depend on the exact formula used for the computation of the electromagnetic radiation in each point of the network, achieving good trade-offs between charging efficiency and radiation control; it also exhibits good energy balance properties. We provide extensive simulation results supporting our claims and theoretical results.

2 Related Work

Electromagnetic radiation. The topic of radiation impact has attracted the attention of researchers in many different fields. In that respect, we mention the book of [22], which focuses on reliability and radiation effects in compound semiconductors. However, the radiation aspect remains largely unexplored in the context of wireless networking, at least as far a distributed computing perspective is concerned.

We point out that known adaptive power control methods in cognitive networks do not focus on the aspect of radiation impact, and although some of such methods can be somehow applied to the radiation context, we believe that novel models and methodologies are needed to fully address the particular challenges created. Also, we note that, although the problem is related to energy optimization, in the rich state of the art for energy management in wireless networks (see e.g., [20, 27, 44]), a combined energy/radiation approach as proposed in this paper is missing.

The problem of scheduling wireless transmissions under signal to interference-plus-noise ratio (SINR) constraints, nicely studied in, e.g., [11, 17, 24], may look similar somehow but is different. Protocols handling the interference problem in wireless transmissions mainly focus on correct data transmission and reception; they usually run on the MAC layer of the network stack and some of their techniques include signal coding, time division multiplexing, and back-off schemes. On the other hand, in the Minimum Radiation Path problem, we focus on the total radiation emitted during data propagation (and particularly during simultaneous transmissions) handled at a higher network layer, which gets received by an entity moving along a path inside the network area.

We note that the problem of distributed navigation in WSNs as well as finding minimum exposure paths have a similar flavor to the problem presented in this chapter, but our approach is different. Traditional solutions to navigation problems rely on flooding either the whole network or only smaller parts of it. The algorithms presented in this chapter are online and do not require the existence of an initialization phase. As a result, the communication cost is smaller and it is much faster to adapt to changes in the network.

Also, coverage aspects studied, e.g., in [13] also relate to radiation but their contribution to radiation is indirect. Furthermore, [31] nicely minimizes energy at the routing layer but radiation aspects are not taking into account. Also, geometric versions of coverage problems, such as in [12] are relevant, but again the treatment of radiation needs a more direct approach. Furthermore, in [5], the authors suggest exploiting mobility (especially when it is high) to lower transmission power; this indirectly reduces radiation too. In a similar line, [39] studies the dynamics of information dissemination between agents moving independently on a plane, providing tight bounds. In [25], the authors suggest methods for optimizing broadcasting in radio networks. Another similar yet different problem is that of synchronization (see e.g., [9]). Although these nice methods can indirectly contribute to radiation reduction, our approach addresses radiation explicitly.

Finally, in [4], we continue our research from [35] by studying the problem of minimizing radiation levels across the network area during multiple transmissions on multiple paths, i.e.,we address the problem of radiation aware data propagation from the network operation perspective.

Wireless energy transfer and WDS. The technology of highly efficient Wireless Energy Transfer (WET) was proposed for efficient energy transmission over mid-range distances. The work in [26] has shown that through strongly coupled magnetic resonances, the efficiency of transferring 60 W of power over a distance in excess of 2 m is as high as 40%. Industry research also demonstrated that it is possible to improve transferring 60 W of power over a distance of up to 1 m with efficiency of 75% [19]. At present, commercial products utilizing Wireless Energy Transfer have been available on the market such as those in [32, 40, 41]. Finally, the Wireless Power Consortium [42], with members including IC manufacturers, smartphone makers, and telecom operators, was established to set the international standards for interoperable wireless charging.

Research efforts in WDS have already started considering network models that take into account WET technologies. For instance, wireless rechargeable sensor networks consist of sensor nodes, as well as few nodes with high energy supplies (wireless chargers). The latter are capable of fast charging sensor nodes, using Wireless Energy Transfer technologies. In [2, 3], the authors assume a single special mobile charging entity, which traverses the network and wirelessly replenishes the energy of sensor nodes. Their methods are distributed, adaptive, use limited network information and perform well in detailed experimental simulations. In [29, 30], the authors employ multiple mobile chargers in sensor networks and collaboratively compute the coordination, trajectory, and charging processes. The authors also provide protocols that grant the chargers the ability to charge each other. In [34], the authors propose protocols that focus on charging efficiency and energy balance and they perform the evaluation through an experimental setting of real WET devices. In [28], a practical and efficient joint routing and charging scheme are proposed. In [21], the authors consider the problem of scheduling mobile chargers in an on-demand way to maximize the covering utility, the authors formulate the scheduling problem as an optimization one and the authors provide three heuristics. In [43], the authors formulate a set of power flow problems and propose algorithms to solve them based on a detailed analysis on the problem structure. Moreover, the authors further investigate the joint data and power flow problems. In [16], the authors propose a framework of joint wireless energy replenishment and anchor-point-based mobile data gathering in sensor networks by considering various sources of energy consumption and time-varying nature of energy replenishment.

The attention of researchers from many diverse research fields has been drawn to the field of electromagnetic radiation impact. Consequently, there has also been research on radiation-related problems in the WDS context. In [33], the authors study the problem of electromagnetic radiation in wireless sensor networks and more specifically maintaining low radiation trajectories for a person moving in a sensor network area. The authors evaluate, mathematically, the radiation in well-known sensor network topologies and random geometric graphs. Then, the authors implement online protocols and comparatively study their performance via simulation. Those heuristics achieve low radiation paths which are even close to an offline optimum. In [1], the authors focus on the problem of efficient data propagation in wireless sensor networks, trying to keep latency low while maintaining at low levels the radiation cumulated by wireless transmissions. The authors first propose greedy and oblivious routing heuristics that are radiation aware. They then combine them with temporal back-off schemes that use local properties of the network in order to spread radiation in a spatiotemporal way. The proposed radiation-aware routing heuristics succeed to keep radiation levels low, while not increasing latency. In [23], the authors consider the problem of covering a planar region, which includes a collection of buildings, with a minimum number of stations so that every point in the region is within the reach of a station, while at the same time no building is within the dangerous range of a station. However, those approaches are oriented toward network devices radiation, not addressing wireless chargers.

Some limited research has also been conducted in the cross-section of Wireless Energy Transfer and electromagnetic radiation in networking settings. In [6], the authors study the problem of scheduling stationary chargers so that more energy can be received while no location in the field has electromagnetic radiation (EMR) exceeding a given threshold. The authors design a method that transfers the problem to two traditional problems, namely, a multidimensional 0/1 knapsack problem and a Fermat–Weber problem. The method includes constraint conversion and reduction, bounded EMR function approximation, area discretization and expansion, and a tailored Fermat–Weber algorithm. In order to evaluate the performance of their method, the authors build a testbed composed of eight chargers. In [8], the authors consider the problem of scheduling stationary chargers with adjustable power, namely how to adjust the power of chargers so as to maximize the charging utility of the devices, while assuring that EMR intensity at any location in the field does not exceed a given threshold. the authors present an area discretization technique to help re-formulating the problem into a traditional linear programming problem. Further, the authors propose a distributed redundant constraint reduction scheme to cut down the number of constraints, and thus reduce the computational efforts of the problem. Although thematically [8] is related to our current work, nevertheless, our treatment of the subject of LREC is radically different. Indeed, this is due to the different charging model that we define, which takes into account hardware restrictions of the chargers and nodes of the system (energy and capacity bounds). These constraints introduce a nonlinearity in our problems that did not appear in the treatment of [8].

3 Radiation Awareness in Three-Dimensional Wireless Sensor Networks

In this section, we present our research related to finding minimum radiation paths in a WDS [35]. More specifically, in this setting

(a) we first evaluate, both mathematically and by simulation, the radiation in well-known sensor network topologies, such as the grid topology, the random geometric graph (Sects. 3.2 and 3.3).
(b) we then focus on the MRP problem of finding low radiation trajectories for a person moving in a sensor network (see Definition 3), i.e., the person moves from a starting point A of the network to a destination point B by following a minimum radiation path. More specifically, we identify the (offline) optimum path given by a linear program that we provide and we present three online approaches: (i) the minimum distance approach, (ii) the heuristic that greedily minimizes *the next step radiation*, and (iii) a trade-off between minimizing the distance and minimizing *the radiation that we expect*. In each of these methods (presented in more detail in Sect. 3.4), we make the minimal assumptions that the entity moving has knowledge of the target location B and can compute the expected radiation at any point at distance at most r from its current position (Sect. 3.4).

(c) we implement the proposed protocols and comparatively study their performance via simulation; the evaluation shows that our heuristic achieve low radiation paths which are even near-optimal, as the comparison with an off-line optimum (via a LP which we provide) suggests (Sects. 3.5 and 3.6).
(d) via detailed experiments, we investigate the impact of diverse mobility levels to the performance of our heuristics showing that, interestingly, high mobility favors the naive, minimum distance approach, while the MinDRD heuristic works quite well for low mobility, being able to more accurately predict the future state of the network (Sect. 3.6.2).

3.1 Network Model and Radiation Definition

Suppose that we deploy n sensors in a three-dimensional target area $\mathscr{A} \subset \mathbb{R}^3$. Ideally, each sensor corresponds to a single point inside $\mathscr{A}$. For two points $\mathbf{x}, \mathbf{y} \in \mathscr{A}$, we denote by $dist(\mathbf{x}, \mathbf{y})$ the Euclidean distance between the two points. Let, also, r be the transmission range of the sensors. Assume now that each sensor $\mathbf{v}$ samples the environment according to a Poisson process with parameter λ and generates data obtained by this sampling, to be broadcast to its immediate neighborhood. The corresponding Poisson processes (for sampling) for distinct sensors are independent (however, the data of nearby sensors may be correlated). When a sensor detects an event e, it decides to make a transmission of data. The duration of such a transmission is assumed to be an exponential random variable T_e with parameter λ', for some λ' depending on the data packet size and the environment (i.e., data may be retransmitted due to noise and/or collisions). We also assume that the transmission duration is independent of the event generation process. We will also denote by $time(e)$ the time of occurrence of e (with respect to some initial time 0).

3.1.1 Point Radiation

We give the following definition for point radiation:

Definition 1 (*Total radiation during a time interval*) We define the radiation at point $\mathbf{x} \in \mathscr{A}$ caused by sensor $\mathbf{v}$ because of data transmissions related to an event e to be

$$R_{\mathbf{x},e,\mathbf{v}} = B\frac{r^2}{(1 + dist(\mathbf{x}, \mathbf{v}))^2}T_e, \tag{1}$$

where B is a constant depending on the environment.[1]

[1] For example, B may depend on the presence of obstacles in a smart building or the type of the human body part (different types of tissue, bones, etc.). The constant B also captures the linear relation of EMR with the received power. More details, as well as the relevant experimental verification can be found in [7].

We also define the total radiation caused at point $\mathbf{x} \in \mathscr{A}$ from data transmissions due to events occurring in $[t_1, t_2]$ to be

$$R_{\mathbf{x}}([t_1, t_2]) = \sum_{\mathbf{v}} \sum_{e:t_1 \leq time(e) \leq t_2} R_{\mathbf{x},e,\mathbf{v}}. \tag{2}$$

Notice that $R_{\mathbf{x}}([t_1, t_2])$ is a random variable depending on the event interarrival times in the Poisson processes corresponding to each sensor in the network, as well as the uniform random variables of the transmission durations.

We also give the following definition for maximum radiation:

Definition 2 (*Maximum radiation in a time interval*) Given a small time distance $\tau > 0$, the maximum radiation at $\mathbf{x}$ within $[t_1, t_2]$ is the random variable

$$MaxR_{\mathbf{x}}([t_1, t_2], \tau) = \max_{t_1 \leq t \leq t_2 - \tau} R_{\mathbf{x}}([t, t + \tau]). \tag{3}$$

Note here that the time interval τ as well as the rate λ for the data generation is, in general, quite larger (i.e., measured in seconds, minutes, or hours) than the mean transmission duration $\frac{1}{\lambda}$. Therefore, the radiation $R_{\mathbf{x}}([t, t + \tau])$ can be very close to the actual radiation within $[t, t + \tau]$.

3.1.2 Path Radiation

Let $\mathscr{P}$ be a specific (finite) trajectory inside $\mathscr{A}$. Assume that a particle travels on $\mathscr{P}$ with constant velocity. We denote by $\mathscr{P}[\tau_1, \tau_2]$ the part of $\mathscr{P}$ that the particle traverses between time τ_1 and time τ_2. The radiation that is caused to the particle on $\mathscr{P}$ from events occurring in $[t_1, t_2]$ is given by

$$R_{\mathscr{P}}([t_1, t_2]) = \sum_{\mathbf{v}} \sum_{e:t_1 \leq time(e) \leq t_2} \int_{\mathscr{P}[time(e), time(e)+T_e]} B \frac{r^2}{(1 + dist(\mathbf{x}, \mathbf{v}))^2} d\mathbf{x}, \tag{4}$$

Instead of computing the above integral, we can approximate it by the following summation: Let dt be a very small time interval and let $\mathbf{x}_i$ be the position of the particle at time $t_1 + idt$. Then,

$$R_{\mathscr{P}}([t_1, t_2]) \approx \sum_{\mathbf{v}} \sum_{e:t_1 \leq time(e) \leq t_2} \sum_{i:t_1 + idt \in [time(e), time(e)+T_e]} B \frac{r^2}{(1 + dist(\mathbf{x}_i, \mathbf{v}))^2} dt. \tag{5}$$

It is evident that the smaller dt is, the better the above approximation is for the actual value of $R_{\mathscr{P}}([t_1, t_2])$.

Given a small time distance $\tau > 0$, we also define the maximum radiation at a particle moving on $\mathscr{P}$ within $[t_1, t_2]$ to be the random variable

$$MaxR_{\mathscr{P}}([t_1, t_2], \tau) = \max_{t_1 \le t \le t_2 - \tau} R_{\mathscr{P}}([t, t + \tau]). \tag{6}$$

This definition is relevant in the case where one is interested in the maximum radiation that an entity can be exposed to when traversing $\mathscr{P}$ and not the total variation from beginning to end.

3.1.3 Minimum Radiation Paths

We are interested in finding exact or approximate algorithmic solutions to the following problem:

Definition 3 (*Minimum Expected Radiation Path Problem—MRP*) Let G be a WSN deployed in a target area $\mathscr{A}$ and let A, B be two distinct points inside $\mathscr{A}$. Consider an entity moving with constant speed inside $\mathscr{A}$. Find a trajectory $\mathscr{P}$ from A to B inside the target area that starts from A, ends at B and minimizes the expected radiation that the entity is exposed to while traversing $\mathscr{P}$.

We stress the fact that the path $\mathscr{P}$ does not necessarily include the nodes of the network. As a matter of fact, since the expected radiation near those points is high, one should try to avoid them when moving from a point A to B. We also note that the experimental evaluation shows that, further to the total path radiation, our heuristics avoid high radiation levels throughout the path (i.e., not only the aggregate radiation is low but also the radiation at all individual path intervals).

3.2 Point Radiation in Random Geometric Graphs

We will assume that the target area $\mathscr{A}$ of the random geometric graphs model $\mathscr{G}_{n,r}$ is $B(\mathbf{x}, r_{\mathscr{A}})$, i.e., the sphere of radius $r_{\mathscr{A}}$ centered at $\mathbf{x}$. The following theorem concerns the mean and variance of the total radiation caused at point $\mathbf{x}$ from data transmissions due to events occurring in some fixed time interval $[t_1, t_2]$.

Theorem 1 *Let $G_{n,r}$ be a random instance of the random geometric graphs model on target area $\mathscr{A} = B(\mathbf{x}, r_{\mathscr{A}})$. The following hold for the total radiation caused at point $\mathbf{x}$ from data transmissions due to events occurring in some fixed time interval $[t_1, t_2]$:*

$$E[R_{\mathbf{x}}([t_1, t_2])] = \frac{3nB\Lambda r^2}{\lambda' r_{\mathscr{A}}^3} \left(\frac{2r_{\mathscr{A}} + r_{\mathscr{A}}^2}{1 + r_{\mathscr{A}}} - 2\log(1 + r_{\mathscr{A}}) \right), \tag{7}$$

and

$$Var[R_{\mathbf{x}}([t_1,t_2])] = \frac{nB^2r^4}{\lambda'^2}\left(\frac{\Lambda^2+2\Lambda}{(1+r_{\mathscr{A}})^3} - \frac{9\Lambda^2}{r_{\mathscr{A}}^6}\left(\frac{2r_{\mathscr{A}}+r_{\mathscr{A}}^2}{1+r_{\mathscr{A}}} - 2\log(1+r_{\mathscr{A}})\right)^2\right), \tag{8}$$

where $\Lambda = \lambda(t_2 - t_1)$.

Proof We will denote by V the set of nodes that are randomly deployed inside $B(\mathbf{x}, r_{\mathscr{A}})$. Let $X_{\mathbf{v}}$ be the random variable that corresponds to the part of the radiation at point $\mathbf{x}$ because of node $\mathbf{v} \in V$. Clearly, $R_{\mathbf{x}}([t_1,t_2]) = \sum_{\mathbf{v}\in V} X_{\mathbf{v}}$. Let also $N_{\mathbf{v}}([t_1,t_2])$ denote the number of events of the Poisson process corresponding to node $\mathbf{v}$ that happen inside $[t_1,t_2]$. By Definition, 1 we have that

$$\begin{aligned} E[R_{\mathbf{x}}([t_1,t_2])] &= E\left[\sum_{\mathbf{v}}\sum_{e:t_1\le time(e)\le t_2} B\frac{r^2}{(1+dist(\mathbf{x},\mathbf{v}))^2}T_e\right] \\ &= nE[N_{\mathbf{v}}([t_1,t_2])]E[T_e]E\left[\frac{Br^2}{(1+dist(\mathbf{x},\mathbf{v}))^2}\right] \\ &= nB\Lambda\frac{1}{\lambda'}E\left[\frac{r^2}{(1+dist(\mathbf{x},\mathbf{v}))^2}\right], \end{aligned} \tag{9}$$

where $dist(\mathbf{x},\mathbf{v})$ is the random variable of the Euclidean distance of a randomly placed node $\mathbf{v}$ from $\mathbf{x}$. The second equality is justified by linearity of expectation. In the third equality, we used Wald's equality, since the random variable $N_{\mathbf{v}}([t_1,t_2])$ is independent to the T_e, by assumption. Furthermore, the mean number of events in a Poisson process with parameter λ inside $[t_1,t_2]$ is $E[N_{\mathbf{v}}([t_1,t_2])] = \lambda(t_2-t_1) = \Lambda$ and also $E[T_e] = \frac{1}{\lambda'}$ since T_e is exponentially distributed with parameter λ'.

Notice now that $\Pr(dist(\mathbf{x},\mathbf{v}) \le \mathbf{y}) = \frac{\mathbf{y}^3}{\mathbf{r}_{\mathscr{A}}^3}$, therefore by (9) we have that

$$\begin{aligned} E[R_{\mathbf{x}}([t_1,t_2])] &= nB\Lambda\frac{1}{\lambda'}\int_0^{r_{\mathscr{A}}}\frac{r^2}{(1+y)^2}\frac{3y^2}{r_{\mathscr{A}}^3}dy \\ &= \frac{3nB\Lambda r^2}{\lambda' r_{\mathscr{A}}^3}\left(\frac{2r_{\mathscr{A}}+r_{\mathscr{A}}^2}{1+r_{\mathscr{A}}} - 2\log(1+r_{\mathscr{A}})\right) \end{aligned}$$

This establishes the first part of the theorem.

For the second of the theorem, notice that, by definition, the random variables $X_{\mathbf{v}}$ are independent. Consequently,

$$Var[R_{\mathbf{x}}([t_1,t_2])] = nVar[X_{\mathbf{v}}] = n\left(E[X_{\mathbf{v}}^2] - E^2[X_{\mathbf{v}}]\right). \tag{10}$$

Since $E[X_{\mathbf{v}}]$ is known from the first part of the proof, we only need to find an exact formula for $E[X_{\mathbf{v}}^2]$.

We now list in increasing order of occurrence the events produced by the Poisson process corresponding to node $\mathbf{v}$ after time t_1, i.e., $e_1, e_2, \ldots$. Setting for the sake of compactness $Y_{e_i,\mathbf{v}} = \frac{T_{e_i}}{(1+dist(\mathbf{x},\mathbf{v}))^2}$, we have that

$$\begin{aligned}
E[X_{\mathbf{v}}^2] &= B^2 r^4 E\left[\left(\sum_{i=1}^{N_{\mathbf{v}}([t_1,t_2])} Y_{e_i,\mathbf{v}}\right)^2\right] \\
&= B^2 r^4 E\left[\sum_{0<i\neq j\leq N_{\mathbf{v}}([t_1,t_2])} Y_{e_i,\mathbf{v}} Y_{e_j,\mathbf{v}} + \sum_{i=1}^{N_{\mathbf{v}}([t_1,t_2])} Y_{e_i,\mathbf{v}}^2\right] \\
&= B^2 r^4 \left(E[N_{\mathbf{v}}([t_1,t_2])(N_{\mathbf{v}}([t_1,t_2])-1)]E[Y_{e_i,\mathbf{v}}Y_{e_j,\mathbf{v}}] + E[N_{\mathbf{v}}([t_1,t_2])]E[Y_{e_i,\mathbf{v}}^2]\right) \\
&= B^2 r^4 \left(\Lambda^2 E[Y_{e_i,\mathbf{v}}Y_{e_j,\mathbf{v}}] + \Lambda E[Y_{e_i,\mathbf{v}}^2]\right). \qquad (11)
\end{aligned}$$

The second to last equation follows from a straightforward generalization of Wald's equality, while the last equation follows from the fact that $N_{\mathbf{v}}([t_1,t_2])$ is distributed according to Poisson distribution with mean value Λ, i.e., $\Pr(N_{\mathbf{v}}([t_1,t_2]) = i) = e^{\Lambda}\frac{\Lambda^i}{i!}$, $i = 0, 1, \ldots$. Furthermore, we note that $e_i \neq e_j$ means that the random variables corresponding to the transmission duration T_{e_i} and T_{e_j} are independent. Since T_e is exponentially distributed with parameter λ', we have that $E[T_e^2] = \frac{2}{\lambda'^2}$, so by independence of $dist(\mathbf{x},\mathbf{v})$ to T_{e_i} and T_{e_j}, (11) becomes

$$\begin{aligned}
E[X_{\mathbf{v}}^2] &= \frac{B^2 r^4(\Lambda^2+2\Lambda)}{\lambda'^2} E\left[\frac{1}{(1+dist(\mathbf{x},\mathbf{v}))^4}\right] \\
&= \frac{B^2 r^4(\Lambda^2+2\Lambda)}{\lambda'^2} \int_0^{r_{\mathscr{A}}} \frac{1}{(1+y)^4}\frac{3y^2}{r_{\mathscr{A}}^3} dy \\
&= \frac{B^2 r^4(\Lambda^2+2\Lambda)}{\lambda'^2} \frac{1}{(1+r_{\mathscr{A}})^3}. \qquad (12)
\end{aligned}$$

By (10), we then have that

$$Var[R_{\mathbf{x}}([t_1,t_2])] = \frac{nB^2 r^4}{\lambda'^2}\left(\frac{\Lambda^2+2\Lambda}{(1+r_{\mathscr{A}})^3} - \frac{9\Lambda^2}{r_{\mathscr{A}}^6}\left(\frac{2r_{\mathscr{A}}+r_{\mathscr{A}}^2}{1+r_{\mathscr{A}}} - 2\log(1+r_{\mathscr{A}})\right)^2\right).$$

This completes the proof of the theorem. □

As a side note, notice that the right-hand side of (7) is always positive for any $r_{\mathscr{A}} > 0$. Furthermore, when $r_{\mathscr{A}} \gg r$, sensors whose sensing area intersects the boundary of $\mathscr{A}$ can be ignored, since their contribution to the radiation caused on $\mathbf{x}$ is very small.

Suppose, also that $r_{\mathscr{A}}$ is large (ideally $r_{\mathscr{A}} \to \infty$). Then, $E[R_{\mathbf{x}}([t_1,t_2])] \sim \frac{3nB\Lambda r^2}{\lambda' r_{\mathscr{A}}^2}$ and $Var[R_{\mathbf{x}}([t_1,t_2])] \sim \frac{nB^2 r^4(\Lambda^2+2\Lambda)}{\lambda'^2}\frac{1}{r_{\mathscr{A}}^3}$, where $\Lambda = \lambda(t_2 - t_1)$. We can see, then, that even though the variance is comparable to the second moment of $R_{\mathbf{x}}([t_1,t_2])$

around 0, it is much smaller (by a factor of $r_{\mathscr{A}}$) than its mean value. This justifies the use of the results in Theorem 1 to give approximations for the path radiation of several heuristics in later sections.

3.3 Point Radiation in Nearest Neighbor Random Graphs

The random proximity graphs model, also known as the *nearest neighbor random graphs model*, is defined below:

Definition 4 (*k-nearest neighbor random graphs*) Suppose that we deploy n points independently and uniformly at random in a target area $\mathscr{A} \subseteq \mathbb{R}^3$. Given an integer $k > 0$, if we connect every point with its k closest points, then we get an instance of the k-nearest neighbor random graphs model $\mathscr{G}_{n,k}$.

In this section, we will assume that the target area $\mathscr{A}$ of the k-nearest neighbor random graphs model $\mathscr{G}_{n,k}$ is $B(\mathbf{x}, r_{\mathscr{A}})$, i.e., the sphere of radius $r_{\mathscr{A}}$ centered at $\mathbf{x}$. We are interested in the mean value of the total radiation caused at point $\mathbf{x}$ from data transmissions due to events occurring in some fixed time interval $[t_1, t_2]$. Notice that, even though the deployment of the sensors in this model is similar to $\mathscr{G}_{n,r}$, the power (or radius) of each sensor is different (since it is as large as its distance to its k-th closest neighbor). By Definition 1, this affects the radiation emanating from each sensor. The following definition will be useful for the discussion below:

Definition 5 (*Effective radius*) Let $G_{n,k}$ be an instance of the k-nearest neighbor random graphs model. For any node $\mathbf{v}$, we will use the term effective radius to describe the distance of $\mathbf{v}$ to its k-closest neighbor.

We will need the following lemma which provides an upper bound for the effective radius of any node that holds with probability that tends to 1 as $n \to \infty$.

Lemma 1 *Let $G_{n,k}$ be a random instance of the k-nearest neighbor random graphs model on target area $\mathscr{A} = B(\mathbf{x}, r_{\mathscr{A}})$. For any $k < n$, the effective radius of every node $\mathbf{v}$ is at most $\xi \stackrel{def}{=} r_{\mathscr{A}} \sqrt[3]{2\frac{k \ln n + \ln^3 n}{n-k}}$ with probability at least $1 - O\left(\frac{1}{n^2}\right)$.*

Proof Let $Y_{\mathbf{v}}(k)$ be the effective radius of $\mathbf{v}$. The probability of the event $\{Y_{\mathbf{v}}(k) > z\}$ is equal to the probability that there are at most $k-1$ nodes (other than $\mathbf{v}$) inside the intersection of the ball $B(\mathbf{v}, z)$ with $\mathscr{A}$. By construction $z \leq 2r_{\mathscr{A}}$, therefore $|B(\mathbf{v}, z) \cap B(\mathbf{x}, r_{\mathscr{A}})| \geq \frac{1}{8}|B(\mathbf{v}, z)| = \frac{1}{6}\pi z^3$, we have that

$$\begin{aligned}\Pr(Y_{\mathbf{v}}(k) > z) &\leq \sum_{i=0}^{k-1} \binom{n-1}{i} \left(1 - \frac{\frac{1}{6}\pi z^3}{|\mathscr{A}|}\right)^{n-1-i} \\ &\leq \sum_{i=0}^{k-1} n^i e^{-(n-1-i)\frac{z^3}{2r_{\mathscr{A}}^3}} \\ &\leq k e^{k\ln n-(n-k)\frac{z^3}{2r_{\mathscr{A}}^3}}. \end{aligned} \tag{13}$$

In the first inequality, we used Boole's inequality together with the fact that the probability that at most i nodes lie inside $B(\mathbf{v}, z) \cap B(\mathbf{x}, r_{\mathscr{A}})$ is at most $\binom{n-1}{i}$ $\left(1 - \frac{B(\mathbf{v},z)\cap B(\mathbf{x},r_{\mathscr{A}})}{|\mathscr{A}|}\right)^{n-1-i}$. For the second inequality, we used the well-known facts that $\binom{n-1}{i} \leq n^i$ and $1 + x \leq e^x$ for all x.

From (13) we, then, have that $\Pr(Y_v(k) > z) = O\left(\frac{1}{n^3}\right)$, for any $z \geq r_{\mathscr{A}} \sqrt[3]{2\frac{k\ln n+\ln^3 n}{n-k}}$, and the lemma follows by Boole's inequality. □

The following theorem concerns the mean value of total radiation caused at point $\mathbf{x}$ from data transmissions due to events occurring in some fixed time interval $[t_1, t_2]$. The symbol $\sim$ means asymptotically equal.

Theorem 2 *Let $G_{n,k}$ be a random instance of the k-nearest neighbor random graphs model on target area $\mathscr{A} = B(\mathbf{x}, r_{\mathscr{A}})$. If $n \to \infty$ and $k = O(n^{1-\epsilon})$, for some fixed $\epsilon > 0$, the following holds for the total radiation caused at point $\mathbf{x}$ from data transmissions due to events occurring in some fixed time interval $[t_1, t_2]$:*

$$E[R_{\mathbf{x}}([t_1, t_2])] \sim \frac{3nB\Lambda}{\lambda' r_{\mathscr{A}}^3} \left(\int_0^{\xi} z^2 f(z) dz\right) \left(\frac{2r_{\mathscr{A}} + r_{\mathscr{A}}^2}{1 + r_{\mathscr{A}}} - 2\log(1 + r_{\mathscr{A}})\right), \tag{14}$$

where $\Lambda = \lambda(t_2 - t_1)$ and $f(x)$ is the probability density function of the effective radius of a node $\mathbf{v} \in B(\mathbf{x}, r_{\mathscr{A}} - \xi)$.

Proof Similarly to the proof of Eq. (9) in Theorem 1, we get that

$$E[R_{\mathbf{x}}([t_1, t_2])] = nB\Lambda \frac{1}{\lambda'} E\left[\frac{Y_{\mathbf{v}}(k)^2}{(1 + dist(\mathbf{x}, \mathbf{v}))^2}\right].$$

where $Y_{\mathbf{v}}(k)$ is the effective radius of $\mathbf{v}$. By Lemma 1 we then have

$$E[R_{\mathbf{x}}([t_1, t_2])] = nB\Lambda \frac{1}{\lambda'} E\left[\frac{Y_{\mathbf{v}}(k)^2}{(1 + dist(\mathbf{x}, \mathbf{v}))^2} \middle| Y_{\mathbf{u}}(k) \leq \xi, \forall \mathbf{u} \in V\right] + O(n^{-1}). \tag{15}$$

By Definition 4, the number X of nodes whose distance from the boundary of $\mathscr{A}$ is less than ξ is a binomial random variable with parameters n and $p = 1 - \left(\frac{r_{\mathscr{A}} - \xi}{r_{\mathscr{A}}}\right)^3$.

Notice that by the assumptions of the theorem and the definition of ξ in Lemma 1, we have that $\xi = o(r_{\mathscr{A}})$, so $p \leq \frac{3\xi}{r_{\mathscr{A}}}$. By Markov's inequality, we then have that, for any $\epsilon' \in (0, \epsilon/4)$

$$\Pr(X \geq n^{1-\epsilon'}) \leq \frac{np}{n^{1-\epsilon'}} = o(1).$$

By (15) we then have that

$$E[R_{\mathbf{x}}([t_1, t_2])] \sim nB\Lambda \frac{1}{\lambda'} E\left[\frac{Y_{\mathbf{v}}(k)^2}{(1 + dist(\mathbf{x}, \mathbf{v}))^2} \Big| \mathscr{E}\right].$$

where $\mathscr{E}$ is the event $\{dist(\mathbf{x}, \mathbf{v}) \leq \mathbf{r}_{\mathscr{A}} - \xi \bigcap \mathbf{Y_u}(\mathbf{k}) \leq \xi, \forall \mathbf{u} \in \mathbf{V}\}$

The difference now is that by the definition of the model, given that $dist(\mathbf{x}, \mathbf{v}) \leq \mathbf{r}_{\mathscr{A}} - \xi$ and $Y_{\mathbf{u}}(k) \leq \xi, \forall \mathbf{u} \in V$, the effective radius of $\mathbf{v}$ is independent to its distance from the center of $\mathscr{A}$. Consequently, we have that

$$E[R_{\mathbf{x}}([t_1, t_2])] \sim nB\Lambda \frac{1}{\lambda'} E[Y_{\mathbf{v}}(k)^2 | \mathscr{E}] E\left[\frac{1}{(1 + dist(\mathbf{x}, \mathbf{v}))^2} \Big| \mathscr{E}\right]. \tag{16}$$

Similarly to the proof of Theorem 1 we also have that

$$\begin{aligned} E\left[\frac{1}{(1 + dist(\mathbf{x}, \mathbf{v}))^2} \Big| \mathscr{E}\right] &= \int_0^{r_{\mathscr{A}} - \xi} \frac{1}{(1+y)^2} \frac{3y^2}{r_{\mathscr{A}}^3} dy \\ &\sim \frac{3}{r_{\mathscr{A}}^3} \left(\frac{2r_{\mathscr{A}} + r_{\mathscr{A}}^2}{1 + r_{\mathscr{A}}} - 2\log(1 + r_{\mathscr{A}})\right). \end{aligned} \tag{17}$$

We then only need a tight formula for $E[Y_{\mathbf{v}}(k)^2 | \mathscr{E}]$. As in the proof of Lemma 1, notice that the event $\{Y_{\mathbf{v}}(k) \leq z\}$ means that there are at least k nodes in $B(\mathbf{v}, z)$, i.e., the sphere centered at $\mathbf{v}$ of radius z. Consequently, since conditioning on the event $\mathscr{E}$ means that $B(\mathbf{v}, z)$ does not intersect with the boundary of $\mathscr{A}$, for any $z \leq r_{\mathscr{A}} - \xi$, we have that

$$\Pr(Y_{\mathbf{v}}(k) \leq z | \mathscr{E}) = \sum_{i=k}^{n-1} \binom{n-1}{i} \left(\frac{z}{r_{\mathscr{A}} - \xi}\right)^{3i} \left(1 - \frac{z^3}{(r_{\mathscr{A}} - \xi)^3}\right)^{n-1-i}. \tag{18}$$

By differentiating, we get $\frac{d \Pr(Y_{\mathbf{v}}(k) \leq z)}{dz} = f_{\mathscr{E}}(z)$, so $E[Y_{\mathbf{v}}(k)^2 | \mathscr{E}] = \int_0^{\xi} z^2 f_{\mathscr{E}}(z) dz$. But if now $f(x)$ is the probability density function of the effective radius of a node $\mathbf{v} \in B(\mathbf{x}, r_{\mathscr{A}} - \xi)$ as stated in the theorem, by Lemma 1 and the observation that $Y_{\mathbf{v}}(k) \leq 2r_{\mathscr{A}}$, we have that $\int_0^{\xi} z^2 f(z) dz = E[Y_{\mathbf{v}}(k)^2 | \mathbf{v} \in B(\mathbf{x}, r_{\mathscr{A}} - \xi)] - \int_{\xi}^{2r_{\mathscr{A}}} z^2 f(z) dz = E[Y_{\mathbf{v}}(k)^2 | \mathscr{E}] + O\left(\frac{r_{\mathscr{A}}^2}{n^2}\right)$. This completes the proof. □

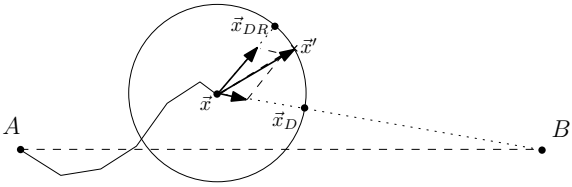


Fig. 1 A dichotomy between MinDR and MinD

Note that it is not obvious how to derive a closed formula for the mean value of the total radiation at point $\mathbf{x}$ during $[t_1, t_2]$ in the case of k-nearest neighbors random graphs model. Indeed, simplifying the expression for $f(z)$ seems far from trivial. Nevertheless, since $f(x)$ can be expressed precisely by differentiating (18), we could use this theorem in a numerical evaluation software environment to compute $E[R_{\mathbf{x}}([t_1, t_2])]$.

3.4 Heuristics for MRP

In this section, we consider several approaches for tackling MRP; the first one simply gives us a "base" solution (moving on the interval AB) with which we can compare the others; the second and third are greedy and online approaches, while the last one is an LP program for the offline optimum (namely the best path $\mathscr{P}$ with respect to radiation between A and B). A visualization of the path constructed by each one of the different heuristics appears in Fig. 2.

3.4.1 Minimizing the Total Distance—Algorithm MinD

This is simply the path defined by the straight line connecting A and B (Fig. 1). It is only intended for comparison purposes. As can be seen by our experiments, this naive approach seems to provide good results in high mobility networks.

3.4.2 Minimizing the Next Step Radiation—Algorithm MinR

We assume that the entity moving has knowledge of the target location B. Furthermore, given that we are at some point $\mathbf{x}$, we assume that the moving entity can compute the radiation at any point located inside $B(\mathbf{x}, r)$, i.e., the ball of radius r centered at $\mathbf{x}$ (which is possible if all the sensors in the larger ball $B(\mathbf{x}, 2r)$ can be heard from it).

At any point $\mathbf{x}$ during its movement, the entity performs the following step: Let $S(\mathbf{x}, r, \phi)$ be the area of all points $\mathbf{x}'$ that satisfy $\left\|\mathbf{x} - \mathbf{x}'\right\| \leq r$ and $\widehat{B\mathbf{x}\mathbf{x}'} \leq \frac{\phi}{2}$ (the angle ϕ can be, e.g., 180°.).

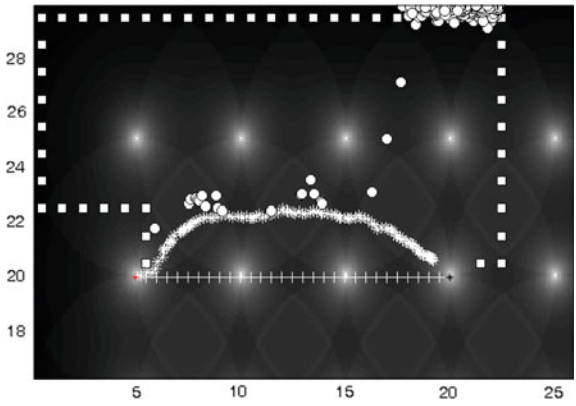


Fig. 2 The path each algorithm follows: In squares is shown the LP's route, in circles the MinR's route, in "*" the MinDRD's route and in "+" the MinD's route

1. If $B \in B(\mathbf{x}, r)$, then move in a straight line to B.
2. Otherwise, choose uniformly at random k points inside $S(\mathbf{x}, r, \phi)$ and move in a straight line to the one that has the lowest expected radiation (k can be, e.g., 3).

The weakness of this approach is that it does not take into account the total distance traveled by the entity and so the resulting path can be quite long. However, this can be reduced by choosing carefully the angle ϕ. More specifically, a small ϕ may reduce the path length, at the cost of increased path radiation. On the other hand, choosing ϕ large may even cause the algorithm to never reach its target. This is illustrated in Fig. 2 in the experiments section.

3.4.3 Optimizing the Radiation/progress to Destination Trade-Off—Algorithm MinDR

In order to overcome the disadvantage of MinR, we present the heuristic MinDR, which takes into account the distance from the target. Even though we do not provide experimental results for this heuristic, we include its description here to facilitate exposition of the MinDRD algorithm in the following section.

We make the same assumptions (on knowledge of B and on the ability to calculate the radiation of points in distance r) as for the MinR heuristic.

At any point $\mathbf{x}$ during its movement, the entity performs the following step:

1. If $B \in B(\mathbf{x}, r)$, then move in a straight line to B.
2. Otherwise, choose uniformly at random k points inside $B(\mathbf{x}, r)$ and move in a straight line to the point $\mathbf{x}'$ that minimizes $R_{\mathbf{y}}([0, 1])\ \|\mathbf{y} - B\|$.

The interval [0, 1] is arbitrary and does not affect the comparison among the k points, since the sampling rates and transmission duration of the sensors are assumed

to be independent and identically distributed. Notice also that this approach takes into account the total distance traveled by the entity, by assuming that once it goes to the new point $\mathbf{x}'$, the radiation levels encountered from that point on will more or less be similar to the expected point radiation at $\mathbf{x}'$. Furthermore, when choosing the candidate points, it does not exclude points that tend to increase the distance from its target. As a matter of fact, it might even select one of these points if its expected radiation level is sufficiently low.

3.4.4 A Dichotomy Algorithm—Algorithm MinDRD

This algorithm is in fact a composition of algorithms MinD and MinDR, since given that we are at some point $\mathbf{x}$, the algorithm considers both moves that MinD and MinDR propose and finally makes a move that is a combination of those moves according to a parameter τ that describes its trust in them. Initially, we set $\tau = 1$. At every time step t, the algorithm does the following:

1. Let $\mathbf{x}_{DR}$ be the point that is suggested by MinDR, let $d = \|\mathbf{x}_{DR} - \mathbf{x}\|$, and let $\mathbf{x}_D$ be the point proposed by MinD (i.e., in the direction of the vector $B - \mathbf{x}$) in euclidean distance d from $\mathbf{x}$.
2. The algorithm computes $\mathbf{x}' = \mathbf{x} + \frac{d(\tau(\mathbf{x}_{DR}-\mathbf{x})+(1-\tau)(\mathbf{x}_D-\mathbf{x}))}{\|\tau(\mathbf{x}_{DR}-\mathbf{x})+(1-\tau)(\mathbf{x}_D-\mathbf{x})\|}$ as the next point. Notice that this is exactly the point in the circle of center $\mathbf{x}$ and radius d, in the direction of the vector $\tau(\mathbf{x}_{DR} - \mathbf{x}) + (1 - \tau)(\mathbf{x}_D - \mathbf{x})$, which is the weighted sum of the suggestions of algorithms MinDR and MinD.
3. Finally, for $s > 0$, let t_s equal to the time needed for an entity of constant speed to travel at distance s (i.e., $t_s = \frac{s}{speed}$). The algorithm updates the parameter of trust to MinDR by

$$\tau' = \alpha\tau + (1 - \alpha) \min \left\{ 1, \frac{R_{\mathbf{x}}([0, t_d])}{\frac{d(R_{\mathbf{x}}([0,t_d])+R_{\mathbf{x}'}([0,t_{dist(\mathbf{x}',B)}]))}{d+dist(\mathbf{x}',B)}} \right\},$$

where α is a fixed parameter of the algorithm that we call *momentum*. Notice that $R_{\mathbf{x}}([0, t_d])$ is the amount of radiation that MinDR expects to have left behind after moving for length d away from $\mathbf{x}$. Similarly, $\frac{d}{d+dist(\mathbf{x}',B)}(R_{\mathbf{x}}([0, t_d]) + R_{\mathbf{x}'}([0, t_{dist(\mathbf{x}',B)}]))$ is the amount of radiation that MinDR expects to have left behind after moving for length d in the path $\mathbf{x} \to \mathbf{x}' \to B$ (i.e., after moving to $\mathbf{x}'$).

The main idea behind MinDRD is that, at any point, it proposes a move according to its level of trust to the MinDR heuristic (i.e., the "belief" that once it goes to a new point $\mathbf{x}'$, the radiation levels encountered from that point on will more or less be similar to the expected point radiation at $\mathbf{x}'$). More specifically, a "good" move, which verifies that the assumption of MinDR is correct, results in strengthening the trust to MinDR (measured by the parameter τ). On the other hand, a "bad" move (i.e.,

a move that increases the cumulative radiation more than MinDR expects) weakens this trust.

3.5 A Linear Program for the Offline Optimum

Without loss of generality, assume that the target area $\mathscr{A}$ is a cube. We tessellate $\mathscr{A}$ in n^3 equal smaller cubes that are characterized by their relative position in the tessellation, with cube in position (1, 1, 1) being the top-left-front one. Furthermore, each cube is represented by its center.

We now construct the following directed graph $G_{n,\mathscr{A}} = (V_{n,\mathscr{A}}, E_{n,\mathscr{A}})$ as follows: The vertex set of $G_{n,\mathscr{A}}$ is the set of all cube center points $\mathbf{v}_{i,j,k}$, namely $V_{n,\mathscr{A}} = \{\mathbf{v}_{i,j,k} : \mathbf{v}_{i,j,k}$ is the central point of the square in position (i, j, k), $1 \leq i, j, k \leq n\}$.

The edge set of $G_{n,\mathscr{A}}$ contains all arcs from any point $\mathbf{v}_{i,j,k}$ to the center points of its neighboring cubes (we say that a cube is a neighbor to another if they share a side). For example, $(\mathbf{v}_{i,j,k}, \mathbf{v}_{i\pm1,j,k}) \in E_{n,\mathscr{A}}$, for all $1 \leq j, k \leq n$ and $2 \leq i \leq n-1$. It is also straightforward to verify that the maximum degree in $G_{n,\mathscr{A}}$ is 6.

Let, now T be the $|V_{n,\mathscr{A}}| \times |E_{n,\mathscr{A}}|$ node-arc adjacency matrix of $G_{n,\mathscr{A}}$. Each row of T corresponds to a different vertex in $V_{n,\mathscr{A}}$ and each column corresponds to a different arc in $E_{n,\mathscr{A}}$. That is, for any $\mathbf{v} \in V_{n,\mathscr{A}}$ and $e = (e_1, e_2) \in E_{n,\mathscr{A}}$, we have

$$T_{v,e} = \begin{cases} 1 & \text{, if } \mathbf{v} = e_1 \\ -1 & \text{, if } \mathbf{v} = e_2 \\ 0 & \text{, otherwise.} \end{cases}$$

A description of our linear program follows: We assume that any path between points A and B is composed by intervals between the points in the center of neighboring cubes of the tessellation. By making the tessellation finer, we get better approximations of the actual path. Let d be the Euclidean distance between the centers of any two neighboring cubes and let t_d be the time needed for an entity of constant speed to travel at distance d (i.e., $t_d = \frac{d}{speed}$). Our problem of finding an optimal path between points A and B can then be reduced to the problem of finding a minimum weight path between vertices A and B in $G_{n,\mathscr{A}}$, assuming that the weight of each edge $e = (e_1, e_2)$ is equal to $w(e) = R_{e_1}([0, t_d])$. Clearly, any path between points A and B is a collection of arcs. For any arc $e \in E_{n,\mathscr{A}}$, let x_e be the indicator variable that this arc is used. Of course, in order for a collection $\mathscr{P}$ of arcs to define a walk between A and B, one needs to guarantee that for any vertex $\mathbf{v} \neq A, B$ in the walk, the number of outgoing arcs is equal to the number of incoming, which is equivalent to $(T\mathbf{x})_{\mathbf{v}} = 0$, where $\mathbf{x} = [x_e]_{e \in E_{n,\mathscr{A}}}$. Also, for A (respectively B) the number of outgoing arcs should be one more (respectively one less) than the number of incoming arcs. Finally, minimizing over all $\mathbf{x}$ that specify a walk, will guarantee that the solution to our LP will be a path.

The corresponding linear program looks like the following:

$$\begin{aligned} \min \quad & \sum_{e \in E_{n,\mathscr{A}}} w(e)x_e \qquad (19) \\ s.t. \quad & \\ & \sum_{e \in E_{n,\mathscr{A}}} T_{A,e}x_e = 1 \\ & \sum_{e \in E_{n,\mathscr{A}}} T_{B,e}x_e = -1 \\ & \sum_{e \in E_{n,\mathscr{A}}} T_{v,e}x_e = 0, \quad \text{for every } \mathbf{v} \in V_{n,\mathscr{A}} \setminus \{A, B\} \\ & x_e \in \{0, 1\}, e \in E_{n,\mathscr{A}}. \end{aligned}$$

Note of course that Dijkstra's Algorithm can be used to find a desired shortest (weighted) path in $G_{n,\mathscr{A}}$ in $O\left(|V_{n,\mathscr{A}}|^2\right) = O(n^6)$ running time.

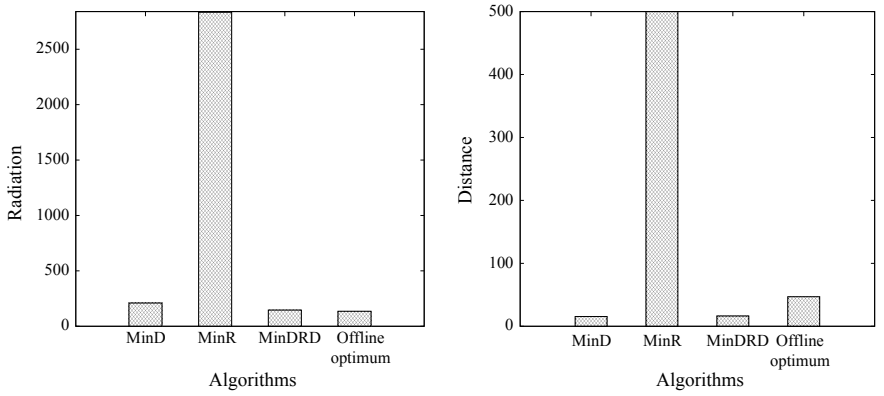
3.6 Performance Evaluation

Simulation Setup. We used Matlab R2008b as our simulation environment. We evaluated the three heuristics and the linear program in a $30\,\text{m} \times 30\,\text{m}$ network region and used two different network topologies for the deployment of the sensor nodes: the grid and the random uniform placement. In both cases, the transmission range of the sensors is set to $R = 5m$. In the grid topology, 25 sensor nodes are used, while for the random uniform deployment we used 100 sensor nodes.

For the linear program evaluation, we tessellate the network area in 900 equal squares. Furthermore, in order to measure the radiation of the MinD heuristic, we break the path in equal parts of size 0.5. For the MinR and MinDRD algorithms, we used an angle $\phi = 180°$. The coordinates of the source and the destination of the path are (5, 20) and (20, 20), respectively.

For each network topology, we conducted 100 iterations and we measured the mean values of the radiation each heuristic resulted to and the distance traveled by the particle (i.e., the trajectory length). The statistical analysis of the findings (the median, lower and upper quartiles, outliers of the samples) demonstrates very high concentration (more than 95%) around their mean value, so in the following figures we only depict average values.

In order to give an intuition for each of the proposed heuristics' behavior, in Fig. 2, we present the path formed by each algorithm, in the grid topology. On the background, one can see the radiation levels at each point of the network area. The brighter the color, the higher the radiation level, with the brightest spots being the radiation at the sensor nodes' locations.



(a) Radiation (b) Distance

Fig. 3 Radiation (**a**) and distance (**b**) of paths from the 4 algorithms in the grid

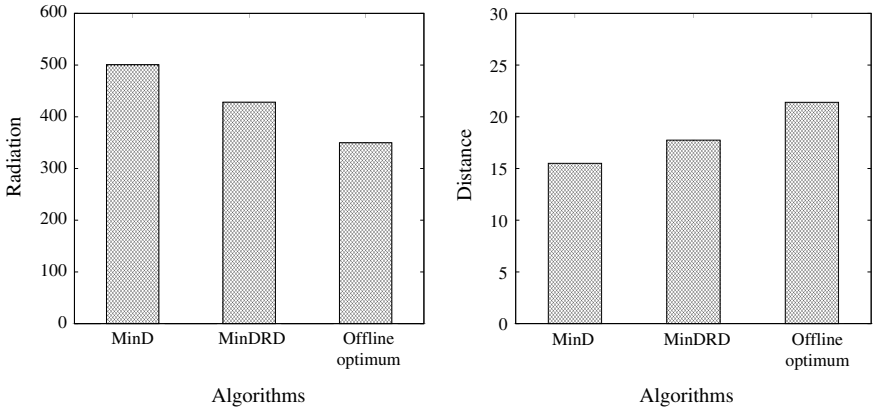
3.6.1 Findings in the Static Scenario

The performance in terms of radiation of the different algorithms in the grid topology is depicted in the left diagram of Fig. 3, while the distance traveled in each algorithm is shown in right diagram of Fig. 3. We observe that the linear program gives the best solution in terms of radiation but the moving entity has to cover longer distance than MinD or MinDRD. Furthermore, the MinDRD algorithm achieves a nice trade-off between total path radiation and distance traveled, since the radiation levels are close to the ones that the linear programs yields and the distance is close to the Euclidean distance between the starting point and the end of the path.

On the other side, the MinR algorithm's performance is poor. This happens because the algorithm tries to minimize the next step radiation, which can result to only small progress towards the target; thus, both the distance traveled and the total radiation of the resulting path increase significantly. Because of this (and also to enhance exposition of the results for the other heuristics), in what follows we avoid presenting other results for MinR.

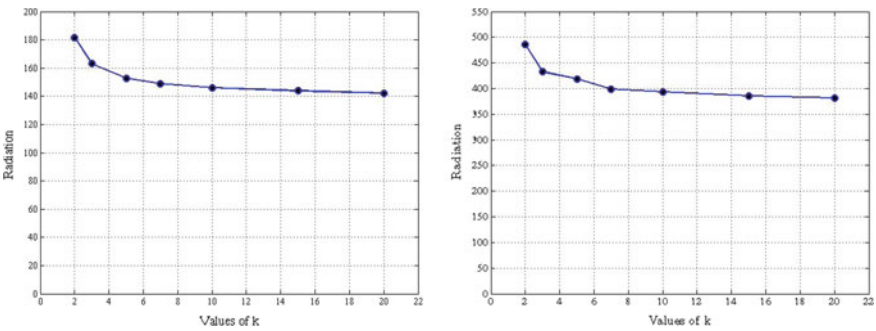
The performance in terms of radiation of the different algorithms in the random uniform placement is depicted in Fig. 4a, while the distance traveled in each algorithm is shown in Fig. 4b. One can see that the algorithms' behavior is similar to their behavior in the grid topology. That is, the linear program gives the offline optimum for the total radiation of the path, while the MinDRD algorithm performs better compared to the MinD algorithm. Moreover, the distance of MinDRD is again very close to the MinD distance. Furthermore, the disadvantage of MinR (compared to the other heuristics) remains.

Multiple samples. Figure 5 presents the total radiation of the path produced by MinDRD using different values of the parameter k, namely the number of the randomly chosen points in step 2 of the MinDR heuristic (which is invoked by MinDRD, see



(a) Radiation (b) Distance

Fig. 4 Radiation (**a**) and distance (**b**) of paths from the 4 algorithms in the random uniform placement

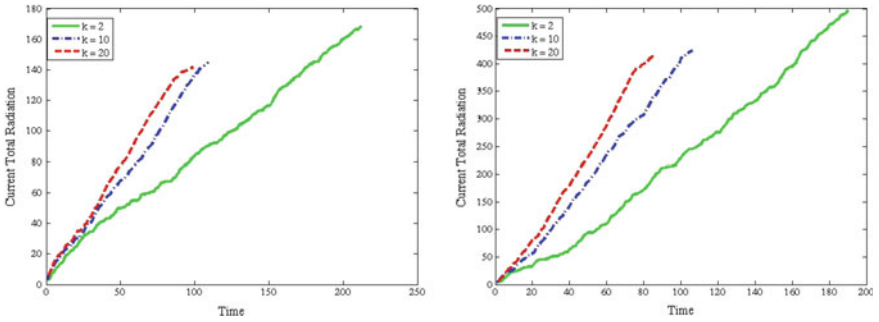


(a) Grid (b) Random Uniform

Fig. 5 Total radiation of MinDRD for different values of the parameter k in the grid (**a**) and the random uniform placement (**b**)

Sects. 3.4.3 and 3.4.4). One can see that as the value of k increases, the radiation tends to decrease. However, the decrement is negligible for $k > 10$. This suggests how a desired compromise between performance and cost can be achieved by accordingly choosing k.

Radiation at path intervals. In order to verify that there are no intervals during the progress of MinDRD where the radiation level is much higher than in the rest of the path, we measured the radiation evolution of the proposed path in time. As can be seen in Fig. 6, the cumulative radiation as a function of time (namely of the number of steps) does not have any peaks.



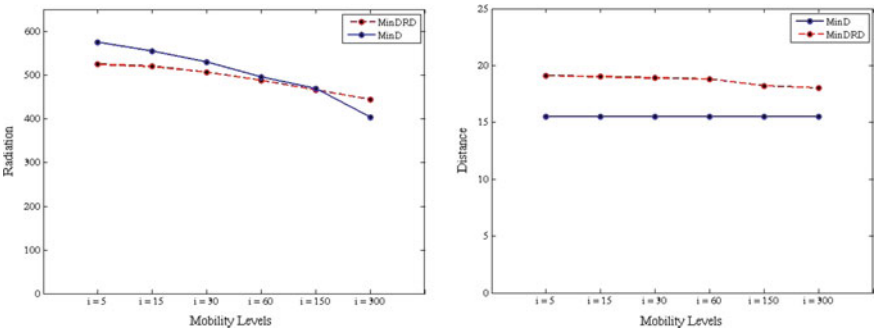
(a) Grid (b) Random Uniform

Fig. 6 Evolution of the radiation level during the execution of MinDRD in the grid (**a**) and random uniform placement (**b**)

3.6.2 Findings in the Mobile Scenario

We use the following method to simulate movement in our network: We define a counter T that keeps track of the number of steps that an algorithm performs. For some positive integer i that is given as input to our program, whenever $T \bmod i = 0$, every node in the network selects uniformly at random a movement direction and then moves towards that direction for distance chosen independently uniformly at random from an arbitrary interval that is also given as input to our program. Therefore, the network changes every i steps of our algorithms. Intuitively, smaller values for i correspond to higher mobility of the nodes and vice versa.

In Fig. 7, we present the total radiation and the length of the path produced by the MinD and MinDRD heuristics for different mobility levels. For MinDRD, we use $k = 10$ for the number of random points examined, which is suggested by our



(a) Radiation (b) Distance

Fig. 7 Radiation (**a**) and distance (**b**) of MinD and MinDRD in a mobile scenario; smaller values for i correspond to higher mobility of the nodes and vice versa

experiments in the static case (see Fig. 5b).It is remarkable that when mobility is high (i.e., i is small), the naive minimum distance heuristic outperforms (in terms of total path radiation) the MinDRD heuristic. However, the difference between the two heuristics becomes smaller as i increases. For values of i larger than a critical value (around $i = 150$), we can see that the MinDRD heuristic outperforms the MinD. The reason for this threshold behavior is because MinDRD uses neighborhood radiation levels to determine the next move. When mobility is low, one can better "predict" (and consequently use) the neighborhood radiation levels, while for high mobility any information on neighborhood radiation levels becomes quickly outdated.

4 Low Radiation Efficient Wireless Energy Transfer in Wireless Distributed Systems

In this section, we present our research related to the problem of efficiently charging a set of rechargeable nodes using a set of wireless energy chargers, under safety constraints on the electromagnetic radiation incurred [36]. In particular, we define a *charging model* that greatly differs from existing models in that it takes into account hardware restrictions of the chargers and nodes of the system. More precisely, we assume (a) that chargers have *finite initial energy supplies*, which restricts the amount of energy that they can transfer to nearby nodes and (b) that every node has *finite battery capacity*, which restricts the total amount of energy that it can store. It is worth noting that previous works have only considered the problem of maximization of the energy transfer rate from the chargers to the nodes, thus ignoring such restrictions. However, new technological advances on wireless energy transfer via strongly coupled magnetic Resonances suggest that such restrictions are already in the heart of efficient energy management problems in such systems.

An important consequence of the energy and capacity restrictions in our model, which sets it apart from other models considered in the literature thus far, is that they introduce *nonlinear constraints* that radically change the nature of the computational problems we consider. In fact, our charging model implicitly introduces the notion of *activity time* in the (radiation aware) charging process, which is the time that a wireless entity (i.e., charger or node) can "affect the system.

In this charging model, we present and study the LREC Problem. The objective function that we wish to optimize in LREC is the amount of "useful" energy transferred from chargers to nodes (under constraints on the maximum level of radiation caused because of the Wireless Energy Transfer). We present several fundamental properties of our objective function that highlight several obstacles that need to be overcome when studying LREC. Furthermore, we present an algorithm for computing the value of the objective function, given the configuration of the system at any time point, which runs in linear time in the number of chargers and nodes.

Furthermore, we present a relaxation of the LREC problem, namely the *Low Radiation Disjoint Charging Problem* (RLDC), which simplifies the computation of

the maximum electromagnetic radiation inside the area where chargers and nodes are deployed (i.e., the area of interest). We prove that even this seemingly easier version of our basic problem is NP-hard, by reduction from the Independent Set Problem in Disc Contact Graphs. Furthermore, we present an integer program for finding the optimal solution to RLDC. We approximately solve this integer program by using standard relaxation and rounding techniques and we use the computed (feasible) solution to assess the performance of our iterative heuristic solution to LREC.

In view of hardness indications for LREC, we propose an iterative local improvement heuristic `IterativeLREC`, which runs in polynomial time and we evaluate its performance via simulation. The most important feature of our algorithmic solution is that it decouples the computation of the objective function from the computation of the maximum radiation. Furthermore, our algorithmic solution is independent of the exact formula used for the computation of the point electromagnetic radiation. Finally, we provide extensive simulation results supporting our claims and theoretical results. We focus on three network metrics: *charging efficiency*, *maximum radiation* and *energy balance*.

4.1 Network and Charging Model

We assume that there is a set of n rechargeable *nodes* $\mathscr{P} = \{v_1, v_2, \ldots, v_n\}$ and a set of m wireless power *chargers* $\mathscr{M} = \{u_1, u_2, \ldots, u_m\}$, which are deployed inside an area of interest $\mathscr{A}$ (say inside $\mathbb{R}^2$). Unless otherwise stated, we will assume that both nodes and chargers are static, i.e., their positions and operational parameters are specified at time 0 and remain unchanged from that time on.

For each charger $u \in \mathscr{M}$, we denote by $E_u^{(t)}$ the *available energy* of that charger, that it can use to charge nodes within some *radius* r_u (i.e., we assume that the initial energy of charger u is $E_u^{(0)}$). The radius r_u for each charger $u \in \mathscr{M}$, can be chosen by the charger at time 0 and remains unchanged for any subsequent time (hence, the nondependence of r_u from t in the notation). Furthermore, for each node $v \in \mathscr{P}$, we denote by $C_v^{(t)}$ the *remaining energy storage capacity* of the node at time t (i.e., the initial energy storage capacity of node v is $C_v^{(0)}$).

We consider the following well-established *charging model*: a node $v \in \mathscr{P}$ harvests energy from a charger $u \in \mathscr{M}$ with *charging rate* given by

$$P_{v,u}(t) = \begin{cases} \frac{\alpha r_u^2}{(\beta + \mathrm{dist}(v,u))^2}, & \text{if } E_u^{(t)}, C_u^{(t)} > 0, \mathrm{dist}(v,u) \leq r_u \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

α and β are known positive constants determined by the environment and by hardware of the charger and the receiver. In particular, the above equation determines the rate at which a node v harvests energy from any charger u that has v within its range, until the energy of u is depleted or v is fully charged. We stress out here that, besides its dependence on the geographic positions of v and u, the charging rate $P_{v,u}(t)$ is also a

function of time t. Indeed, for any node $v \in \mathscr{P}$ within distance r_u from charger u, it is equal to $\frac{\alpha r_u^2}{(\beta+\text{dist}(v,u))^2}$ in a time interval $[0, t^*_{u,v}]$ and 0 otherwise. By Eq. (20), the time point $t^*_{u,v}$ at which the value of $P_{v,u}(t)$ drops to 0 is the time when either the energy of u is depleted or v is fully charged (which also depends on other chargers that can reach v). Consequently, the exact value of $t^*_{u,v}$ may depend on the whole network (see also, the discussion after Definition 6 in Sect. 5), i.e., the location, radius and initial energy of each charger and the location and initial energy storage capacity of each node. As a matter of fact, it seems that there is no "nice" closed formula for $t^*_{u,v}$. Nevertheless, the value of $t^*_{u,v}$ can be found by using the ideas of Sect. 5.1 and, more specifically, using a trivial modification of Algorithm `ObjectiveValue`.

Another crucial assumption on our charging model (which is also widely accepted by physicists) is that the harvested energy by the nodes is additive. Therefore, the total energy that node v gets within the time interval $[0, T]$ is

$$H_v(T) = \sum_{u \in \mathscr{M}} \int_0^T P_{v,u}(t)dt. \tag{21}$$

One of the consequences of our charging model (and, in particular, Eqs. (20) and (21) above) is that $\sum_{u \in \mathscr{M}} E_u \geq \sum_{v \in \mathscr{P}} H_v(T)$, for any $T > 0$. This means that the total energy harvested by the nodes cannot be larger than the total energy provided by the chargers. As yet, another consequence, we have that $\sum_{v \in \mathscr{P}} C_v \geq \sum_{v \in \mathscr{P}} H_v(T)$, for any $T > 0$, i.e., the total energy harvested by the nodes cannot be larger than the total energy that can be stored by all nodes.

To complete the definition of our model, we will make the assumption that the *electromagnetic radiation (EMR)* at a point x is proportional to the additive power received at that point; notice that this assumption is in line with the definitions given earlier in this chapter. In particular, for any $x \in \mathscr{A}$, the EMR at time t on x is given by

$$R_x(t) = \gamma \sum_{u \in \mathscr{M}} P_{x,u}(t), \tag{22}$$

where γ is a constant that depends on the environment and $P_{x,u}(t)$ is given by Eq. (20). We note that, even though this is the usual assumption concerning electromagnetic radiation, the algorithmic solutions that we propose here could also be applied in the case of more general functions for $R_x(t)$ (as long as some quite general smoothness assumption are satisfied; see also Sect. 5.2). We feel that this is especially important, because the notion of electromagnetic radiation is not completely understood in our days.

We finally note that the existence of an energy (upper) bound for each charger and a capacity bound for each node greatly differentiates our model from other works in the literature. Indeed, not only can chargers decide on the length of their charging radius (a slight variation of which has been proposed in [8]), but once each charger has made its decision, all chargers begin charging nodes within their radius until

either their energy has been depleted, or every node within their radius has already reached its energy storage capacity. Furthermore, this characteristic radically changes the nature of the computational problem that we consider (see Sect. 5).

5 Problem Statement and First Results

In general, we would like to use the chargers as efficiently as possible, but we would also like to keep radiation levels within acceptable levels. In particular, we are interested in the following computational problem:

Definition 6 (*Low Radiation Efficient Charging (LREC)*) Let $\mathscr{M}$ be a set of wireless power chargers and $\mathscr{P}$ be a set of rechargeable nodes, which are deployed inside an area of interest $\mathscr{A}$. Suppose that each charger $u \in \mathscr{M}$ initially has available energy $E_u^{(0)}$, and each node $v \in \mathscr{P}$ has initial energy storage capacity $C_v^{(0)}$. Assign to each charger $u \in \mathscr{M}$ a radius r_u, so that the total usable energy given to the nodes of the network is maximized and the electromagnetic radiation at any point of $\mathscr{A}$ is at most ρ. We assume that all chargers start operating simultaneously at time 0 and charging follows the model described in Sect. 4.1.

Let $\mathbf{r} = (r_u : u \in \mathscr{M}), \mathbf{E}^{(0)} = (E_u^{(0)} : u \in \mathscr{M})$ and $\mathbf{C}^{(0)} = (C_v^{(0)} : v \in \mathscr{P})$. In essence, the *objective function* that we want to maximize in the LREC problem is the following:

$$\begin{aligned} f_{\text{LREC}}\left(\mathbf{r}, \mathbf{E}^{(0)}, \mathbf{C}^{(0)}\right) &\stackrel{def}{=} \sum_{v \in \mathscr{P}} \left(\lim_{t \to \infty} C_v^{(t)}\right) \\ &= \sum_{u \in \mathscr{M}} \left(E_u^{(0)} - \lim_{t \to \infty} E_u^{(t)}\right). \end{aligned} \tag{23}$$

The last equality follows from the fact that we are assuming loss-less energy transfer from the chargers to the nodes (obviously this easily extends to lossy energy transfer, but we do not consider such models in this chapter). In fact, we only need to consider finite values for t, because the energy values $E_u^{(t)}$ will be unchanged after time $t^* \stackrel{def}{=} \max_{v \in \mathscr{P}, u \in \mathscr{M}} t_{u,v}^*$, where $t_{u,v}^*$ is the time point at which the value of $P_{v,u}(t)$ drops to 0 (i.e., is the time when either the energy of u is depleted or v is fully charged). Therefore, $f_{\text{LREC}}\left(\mathbf{r}, \mathbf{E}^{(0)}, \mathbf{C}^{(0)}\right) = \sum_{v \in \mathscr{P}} C_v^{(t)} = \sum_{u \in \mathscr{M}} \left(E_u^{(0)} - E_u^{(t)}\right)$, for any $t \geq t^*$. The following lemma provides an upper bound on the value of t^*, which is independent of the radius choice for each charger.

Lemma 2 *t^* can be at most*

$$T^* = \frac{(\beta + \max_{u \in \mathscr{M}, v \in \mathscr{P}} dist(v, u))^2}{\alpha(\min_{u \in \mathscr{M}, v \in \mathscr{P}} dist(v, u))^2} \max_{u \in \mathscr{M}, v \in \mathscr{P}} \{E_u^{(0)}, C_v^{(0)}\}.$$

Proof Since $t^* \stackrel{def}{=} \max_{v\in\mathscr{P},u\in\mathscr{M}} t^*_{u,v}$, we only need to provide an upper bound on $t^*_{u,v}$. To this end, without loss of generality, we assume that there is a charger u_0 and a node v_0 such that u_0 can reach v_0 (hence, also $r_{u_0} \neq 0$) and $t^* = t^*_{u_0,v_0}$. Furthermore, we need to consider two cases, depending on whether $t^*_{u_0,v_0}$ is equal to (a) the time when the energy of the charger u_0 is depleted, or (b) the time when v_0 is fully charged.

In case (a), by the maximality of $t^*_{u_0,v_0}$, we have that

$$E^{(0)}_{u_0} = \sum_{v\in\mathscr{P}} \int_0^{t^*_{u,v}} P_{v,u}(t)dt \geq \int_0^{t^*_{u_0,v_0}} P_{v_0,u_0}(t)dt$$
$$= t^*_{u_0,v_0} \frac{\alpha r^2_{u_0}}{(\beta + \mathrm{dist}(v_0,u_0))^2}, \tag{24}$$

where in the first and last equality, we used the fact that in case (a) $t^*_{u_0,v_0}$ is the time when the energy of the charger u_0 is depleted (hence, v_0 has not yet exceeded its energy storage capacity).

In case (b), by Eq. (21) and the maximality of $t^*_{u_0,v_0}$, we have that

$$C^{(0)}_{v_0} = H_{v_0}(t^*_{u_0,v_0}) = \sum_{u\in\mathscr{M}} \int_0^{t^*_{u_0,v_0}} P_{v,u}(t)dt$$
$$\geq \int_0^{t^*_{u_0,v_0}} P_{v_0,u_0}(t)dt = t^*_{u_0,v_0} \frac{\alpha r^2_{u_0}}{(\beta + \mathrm{dist}(v_0,u_0))^2}, \tag{25}$$

where in the first and last equality, we used the fact that in case (b) $t^*_{u_0,v_0}$ is the time when v_0 is fully charged (hence the energy of u_0 has not been depleted yet).

By Eqs. (24) and (25), we have that

$$t^*_{u_0,v_0} \leq \max\{E^{(0)}_{u_0}, C^{(0)}_{v_0}\} \frac{(\beta + \mathrm{dist}(v_0,u_0))^2}{\alpha r^2_{u_0}}, \tag{26}$$

which completes the proof. □

It is worth noting here that given $\mathbf{r}, \mathbf{E}^{(0)}$ and $\mathbf{C}^{(0)}$, the exact value of $f_{\mathrm{LREC}}\left(\mathbf{r}, \mathbf{E}^{(0)}, \mathbf{C}^{(0)}\right)$ can be computed by using Algorithm `ObjectiveValue` in Sect. 5.1.

We now prove a Lemma that highlights some of the difficulties that we face when trying to find a solution to LREC. Furthermore, it sets LREC apart from other computational problems studied so far in the literature.

Lemma 3 *Let* $\mathbf{r} = (r_u : u \in \mathscr{M}), \mathbf{E}^{(0)} = (E^{(0)}_u : u \in \mathscr{M})$ *and* $\mathbf{C}^{(0)} = (C^{(0)}_v : v \in \mathscr{P})$. *The objective function* $f_{LREC}\left(\mathbf{r}, \mathbf{E}^{(0)}, \mathbf{C}^{(0)}\right)$ *is not necessarily increasing in* $\mathbf{r}$. *Furthermore, the optimal radius for a charger is not necessarily equal to the distance from some node.*

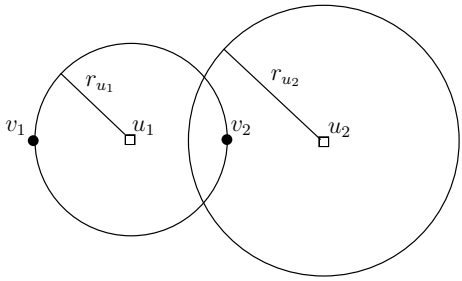


Fig. 8 A network with two chargers u_1, u_2 and 2 nodes v_1, v_2. All four points are collinear and $\text{dist}(v_1, u_1) = \text{dist}(v_2, u_1) = \text{dist}(v_2, u_2) = r_{u_1} = 1$

Proof Consider a network consisting of two chargers u_1, u_2 and 2 nodes v_1, v_2 all of which are collinear and $\text{dist}(v_1, u_1) = \text{dist}(v_2, u_1) = \text{dist}(v_2, u_2) = r_{u_1} = 1$ (see Fig. 8). Furthermore, assume for the sake of exposition of our arguments that $E^{(0)}_{u_1} = E^{(0)}_{u_2} = C^{(0)}_{v_1} = C^{(0)}_{v_2} = 1$. Finally, assume that the parameters for the charging rate in Eq. (20) are $\alpha = \beta = 1$, the electromagnetic radiation parameter in Eq. (22) is $\gamma = 1$ and that the upper bound on the radiation level is $\rho = 2$.

We will show that the optimal solution to the LREC problem in this network is when $r_{u_1} = 1$ and $r_{u_2} = \sqrt{2}$. To see this, first note that the electromagnetic radiation is maximum when $t = 0$, i.e., when chargers are all operational. Furthermore, since there are only two radiation sources (namely chargers u_1 and u_2), it is not hard to verify that the electromagnetic radiation is maximized on the charger locations, i.e., $\max_{x,t} R_x(t) = \max\{r^2_{u_1}, r^2_{u_2}\}$. Consequently, since $\rho = 2$, the radius of each charger can be at most $\sqrt{2}$. On the other hand, to achieve an objective function value that is larger than 1, both r_{u_1} and r_{u_2} must be at least 1. In fact, if $r_{u_1} = r_{u_2} \in [1, \sqrt{2}]$, then, by symmetry, v_2 will reach its energy storage capacity at the exact moment that the energy of u_1 is depleted. Therefore, the objective function value will be only $\frac{3}{2}$, since both u_1 and u_2 will have contributed the same amount of energy to fully charge v_2. To do better, v_2 must reach its energy storage capacity without using too much of the energy of u_1, which happens when $r_2 > r_1$. In this case, v_2 will reach its energy storage capacity before the energy of u_1 is depleted, and so u_1 will use the remaining energy to further charge v_1. In particular, when $r_2 > r_1$, we have that $t^*_{u_2,v_1} = t^*_{u_2,v_2}$ and also

$$1 = C^{(0)}_{v_2} = t^*_{u_2,v_1} \frac{r^2_{u_1} + r^2_{u_2}}{4}. \tag{27}$$

The remaining energy of u_1 at time $t^*_{u_2,v_1}$ will then be

$$E^{(0)}_{u_1} - 2t^*_{u_2,v_1} \frac{r^2_{u_1}}{4} = 1 - 2\frac{r^2_{u_1}}{r^2_{u_1} + r^2_{u_2}}. \tag{28}$$

Since $1 \leq r_1 < r_2 \leq \sqrt{2}$, this is maximized for $r_{u_1} = 1$ and $r_{u_2} = \sqrt{2}$, in which case $E_{u_1}^{(t^*_{u_2,v_1})} = \frac{1}{3}$. In that case, when v_2 reaches its energy storage capacity, u_1 will have $\frac{1}{3}$ energy units more to give solely to v_1; the other $\frac{2}{3}$ units will have been split evenly between v_1 and v_2. This means that the objective function value is $\frac{5}{3}$ and it is maximum.

Now this example shows that not only the radius of chargers are not necessarily equal to the distance from some node (since the to achieve the optimum we must have $r_{u_2} = \sqrt{2}$), but also increasing r_1 will result in a sub-optimal objective function value. □

5.1 Computing the Objective Function

In this section, we provide an algorithm for computing the value of our objective function (i.e., the amount of energy given by the chargers to the nodes), given the radii of the chargers, the capacities of the nodes and the available energies of the chargers. More precisely, assume that at some time t, each charger $u \in \mathscr{M}$ has remaining energy $E_u^{(t)}$ and each node $v \in \mathscr{P}$ can store $C_v^{(t)}$ energy. The tuple $\Sigma^{(t)} = (\mathbf{r}, \mathbf{E}^{(t)}, \mathbf{C}^{(t)})$, where $\mathbf{r} = (r_u : u \in \mathscr{M})$, $\mathbf{E}^{(t)} = (E_u^{(t)} : u \in \mathscr{M})$ and $\mathbf{C}^{(t)} = (C_v^{(t)} : v \in \mathscr{P})$, will be called the *configuration of the system at time t*. For each $u \in \mathscr{M}$, we denote by $\mathscr{P}_u^{(t)} \stackrel{def}{=} \{v : \text{dist}(v, u) \leq r_u, C_v^{(t)} > 0\}$ the set of nodes within distance r_u from u that have not reached their storage capacities at time t. Furthermore, for each $v \in \mathscr{P}$, we denote by $\mathscr{M}_v^{(t)} \stackrel{def}{=} \{u : v \in \mathscr{P}_u^{(t)}, E_u^{(t)} > 0\}$ the set of chargers that can reach v and have not depleted their energy at time t. Finally, denote by $\mathscr{M}_\emptyset^{(t)} \stackrel{def}{=} \{u \in \mathscr{M} : E_u^{(t)} = 0\}$ the set of chargers that have depleted their energy by time t. Similarly, denote by $\mathscr{P}_\emptyset^{(t)} \stackrel{def}{=} \{v \in \mathscr{P} : C_u^{(t)} = 0\}$ the set of nodes that have reached their energy storage capacity by time t.

The value of the objective function can be computed by the following algorithm. The main idea is that given the configuration of the system at any time t, we can find which will be the next charger (or node respectively) that will deplete his energy (respectively will reach its energy storage capacity) and when. The algorithm stops when no node can be charged any more, which happens either when they have reached their total capacity (i.e., $C_v^{(t)} = 0$), or all chargers that can reach it have depleted their energy (i.e., $\sum_{u \in \mathscr{M}_v^{(t)}} E_u^{(t)} = 0$).

Algorithm 1: `ObjectiveValue`

Input : Initial configuration $\Sigma^{(0)} = (\mathbf{r}, \mathbf{E}^{(0)}, \mathbf{C}^{(0)})$

Set $t = 0$

while $\left[\bigcup_{v\in\mathscr{P}} \left\{\left(C_v^{(t)} > 0\right) \; AND \; \left(\sum_{u\in\mathscr{M}_v^{(t)}} E_u^{(t)} > 0\right)\right\}\right]$ **do**

  Let $t_{\mathscr{M}} = \min_{u\in\mathscr{M}\setminus\mathscr{M}_{\emptyset}^{(t)}}\{t' : t'\sum_{v\in\mathscr{P}_u^{(t)}} P_{v,u}(t) = E_u^{(t)}\}$

  Let $t_{\mathscr{P}} = \min_{v\in\mathscr{M}\setminus\mathscr{P}_{\emptyset}^{(t)}}\{t' : t'\sum_{u\in\mathscr{M}_v^{(t)}} P_{v,u}(t) = C_v^{(t)}\}$

  Let $t_0 = \min\{t_{\mathscr{M}}, t_{\mathscr{P}}\}$

  For all $u \in \mathscr{M}\setminus\mathscr{M}_{\emptyset}^{(t)}$, set $E_u^{(t+t_0)} = E_u^{(t)} - t_0\sum_{v\in\mathscr{P}_u^{(t)}} P_{v,u}(t)$

  For all $v \in \mathscr{P}\setminus\mathscr{P}_{\emptyset}^{(t)}$, set $C_v^{(t+t_0)} = C_v^{(t)} - t_0\sum_{u\in\mathscr{M}_v^{(t)}} P_{v,u}(t)$

  Set $t = t + t_0$ and update $\mathscr{M}_{\emptyset}^{(t)}$ and $\mathscr{P}_{\emptyset}^{(t)}$

Output: $\sum_{u\in\mathscr{M}} (E_u^{(0)} - E_u^{(t)})$

Notice that, in every iteration, algorithm `ObjectiveValue` sets to 0 the energy level or the capacity of at least one charger or node. Therefore, we have the following:

Lemma 4 *Algorithm `ObjectiveValue` terminates in at most $n + m$ while-iterations.*

5.2 Computing the Maximum Radiation

One of the challenges that arises in our model is the computation of the maximum radiation inside the area of interest $\mathscr{A}$, as well as the point (or points) where this maximum is achieved. Unfortunately, it is not obvious where the maximum radiation is attained inside our area of interest and it seems that some kind of discretization is necessary. In fact, in our experiments, we use the following generic MCMC procedure: for sufficiently large $K \in \mathbb{N}^+$, choose K points uniformly at random inside $\mathscr{A}$ and return the maximum radiation among those points. We note also that the computation of the electromagnetic radiation at any point takes $O(m)$ time, since it depends only on the distance of that point from each charger in $\mathscr{M}$.

One of the main drawbacks of the above method for computing the maximum radiation is that the approximation it achieves depends on the value of K (which is equivalent to how refined our discretization is). On the other hand, it does not take into account the special form of the electromagnetic radiation in Eq. (22). In fact, our iterative algorithm `IterativeLREC` in Sect. 5.3 does not depend on the specific form of Eq. (22), and this could be desirable in some cases (especially since the effect that multiple radiation sources have on the electromagnetic radiation is not completely understood).

Algorithm 2: `IterativeLREC`

Input : Charger and node locations
1 $counter = 1$
2 **repeat**
3 Select u.a.r. a charger $u \in \mathscr{M}$
4 Find (an approximation to) the optimal radius for u given that the radii of all other chargers are fixed
5 $counter = counter + 1$
6 **until** $counter = K'$
Output: $\mathbf{r} = (r_u : u \in \mathscr{M})$

5.3 A Local Improvement Heuristic for LREC

We now present a heuristic for approximating the optimal solution to LREC. To this end, we first note that for any charger $u \in \mathscr{M}$, we can approximately determine the radius r_u of u that achieves the best objective function value, given the radii $\mathbf{r}_{-u} = (r_{u'} : u' \in \mathscr{M} \setminus u)$ as follows: Let $r_u^{\max}$ be the maximum distance of any point in $\mathscr{A}$ from u and let $l \in \mathbb{N}^+$ be a sufficiently large integer. For $i = 0, 1, \ldots, l$, set $r_u = \frac{i}{l} r_u^{\max}$ and compute the objective function value (using algorithm `ObjectiveValue`) as well as the maximum radiation (using the method described in Sect. 5.2). Assign to u, the radius that achieves the highest objective function value that satisfies the radiation constraints of LREC. Given that the discretization of $\mathscr{A}$ used to compute the maximum radiation has K points in it, and using Lemma 4, we can see that the number of steps needed to approximately determine the radius r_u of u using the above procedure is $O((n+m)l + mK)$. It is worth noting that we could generalize the above procedure to any number c of chargers, in which case the running time would be $O((n+m)l^c + mK)$. In fact, for $c = m$ we would have an exhaustive-search algorithm for LREC, but the running time would be exponential in m, making this solution impractical even for a small number of chargers.

The main idea of our heuristic `IterativeLREC` is the following: in every step, choose a charger u uniformly at random and find (an approximation to) the optimal radius for u given that the radii of all other chargers are fixed. To avoid infinite loops, we stop the algorithm after a predefined number of iterations $K' \in \mathbb{N}^+$.

By the above discussion, `IterativeLREC` terminates in $O\left(K'(nl + ml + mK)\right)$ steps.

5.4 A Relaxation of LREC

The intractability of the LREC problem is mainly due to the following reasons: (a) First, there is no obvious closed formula for the maximum radiation inside the area of interest $\mathscr{A}$ as a function of the positions and the radii of the chargers. (b) Second,

as is suggested by Lemma 3, there is no obvious potential function that can be used to identify directions inside $\mathbb{R}^m$ that can increase the value of our objective function.

In this section, we consider the following relaxation to the LREC problem, which circumvents the problem of finding the maximum radiation caused by multiple sources:

Definition 7 (*Low Radiation Disjoint Charging (LRDC)*) Let $\mathscr{M}$ be a set of wireless power chargers and $\mathscr{P}$ be a set of rechargeable nodes which are deployed inside an area of interest $\mathscr{A}$. Suppose that each charger $u \in \mathscr{M}$ initially has available energy $E_u^{(0)}$, and each node $v \in \mathscr{P}$ has initial energy storage capacity $C_v^{(0)}$. Assign to each charger $u \in \mathscr{M}$ a radius r_u, so that the total usable energy given to the nodes of the network is maximized and the electromagnetic radiation at any point of $\mathscr{A}$ is at most ρ. We assume that all chargers start operating simultaneously at time 0 and that charging follows the model described in Sect. 4.1. Additionally, we impose the constraint that no node should be charged by more than one charger.

The following Theorem concerns the hardness of LRDC.

Theorem 3 *LRDC is NP-hard.*

Proof The hardness follows by reduction from the Independent Set in Disc Contact Graphs [15]. Let G be a disc contact graph, i.e., a graph where vertices correspond to discs any two of which have at most one point in common. In particular, the set of vertices of G corresponds to a set of m discs $D(u_1, r_1), D(u_2, r_2), \ldots, D(u_m, r_m)$, where $D(u_j, r_j)$ is a disc centered at u_j with radius r_j. Two vertices of G are joined by an edge if and only if their corresponding discs have a point in common.

We now construct an instance of the LRDC as follows: We place a node on each disc contact point and, for $j = 1, 2, \ldots, m$, let k_j be the maximum number of nodes in the circumference of the disc $D(u_j, r_j)$. We then add nodes on the circumference of every other disc in such a way that every disc has exactly the same number of nodes (say) K uniformly around its circumference (notice that this is possible since every disc shares at most m points of its circumference with other discs). We now place a charger on the center of each disc and set the radius bound for the charger corresponding to u_j equal to r_j, for every $j = 1, 2, \ldots, m$. Finally, we set the initial energy storage capacity of each node equal to 1, the available energy of each charger equal to K and the electromagnetic radiation bound $\rho = \max_{j \in [m]} \frac{\alpha r_j^2}{\beta^2}$.

It is now evident that an optimal solution to LRDC on the above instance yields a maximum independent set in G; just pick disc $D(u_j, r_j)$ if the j-th charger has radius equal to r_j and discard it otherwise. □

We now present an integer program formulation for LRDC (to which we refer as `IP-LRDC`). To this end, we first note that for any charger $u \in \mathscr{M}$, the distance of nodes/points in $\mathscr{P}$ from u defines a (complete) ordering σ_u in $\mathscr{P}$. In particular, for any two nodes $v, v' \in \mathscr{P}$ and a charger $u \in \mathscr{M}$, we will write $v \preceq_{\sigma_u} v'$ if and only if $\text{dist}(v, u) \leq \text{dist}(v, u)$. For any charger u, define $i_{\text{rad}}^{(u)}$ to be the furthest node from u that can be charged by u without u violating the radiation threshold ρ on its own.

Similarly, define $i_{\text{nrg}}^{(u)}$ to be the furthest node from u with the property that if u has radius at least $\text{dist}(i_{\text{nrg}}^{(u)}, u)$, then the energy of u will be fully spent. Assuming we break in σ arbitrarily, nodes $i_{\text{rad}}^{(u)}$ and $i_{\text{nrg}}^{(u)}$ are uniquely defined for any charger u. Our integer program solution is presented below.

$$\max \sum_{u \in \mathscr{M}} \left(E_u^{(0)} x_{i_{\text{nrg}}^{(u)},u} + \sum_{v \leq_{\sigma_u} i_{\text{nrg}}^{(u)}} (x_{v,u} - x_{i_{\text{nrg}}^{(u)},u}) C_v^{(0)} \right) \tag{29}$$

subject to:

$$\sum_{u \in \mathscr{M}} x_{v,u} \leq 1, \qquad \forall v \in \mathscr{P} \tag{30}$$

$$x_{v,u} - x_{v',u} \geq 0, \quad \forall v, v' \in \mathscr{P}, \forall u \in \mathscr{M} : v \leq_{\sigma_u} v' \tag{31}$$

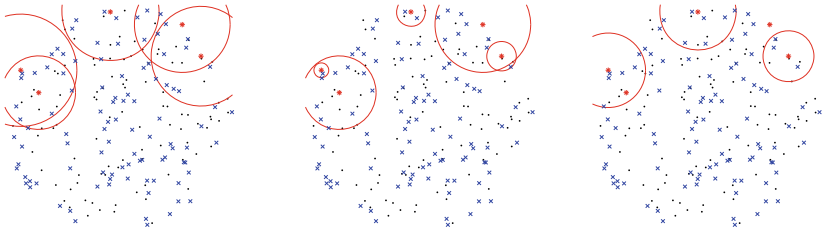
$$x_{v,u} = 0, \quad \forall v \in \mathscr{P}, \forall u \in \mathscr{M} : v >_{\sigma_u} i_{\text{rad}}^{(u)} \text{ or } v >_{\sigma_u} i_{\text{nrg}}^{(u)} \tag{32}$$

$$x_{v,u} \in \{0, 1\}, \qquad \forall v \in \mathscr{P}, \forall u \in \mathscr{M}. \tag{33}$$

In IP-LRDC, variable $x_{v,u}$ indicates whether or not the (unique) charger that can reach v is u. The existence of at most one charger per node in a feasible assignment of LRDC is guaranteed by constraint (30). Constraint (31) guarantees that when a node v' can be reached by u, then all nodes closer to u can also be reached by u. Finally, constraint (32) guarantees that the radiation threshold is not violated and also suggests that there is no reason why a charger should be able to reach nodes that are further than $i_{\text{nrg}}^{(u)}$.

To understand the objective function that we want to maximize in IP-LRDC, notice that, for any charger $u \in \mathscr{M}$, if $r_u \geq \text{dist}(i_{\text{nrg}}^{(u)}, u)$ (which is equivalent to having $x_{i_{\text{nrg}}^{(u)},u} = 1$), then the useful energy transferred from u to the nodes of the network will be exactly $E_u^{(0)}$. Indeed, this is captured by our objective function, since $E_u^{(0)} x_{i_{\text{nrg}}^{(u)},u} + \sum_{v \leq_{\sigma_u} i_{\text{nrg}}^{(u)}} (x_{v,u} - x_{i_{\text{nrg}}^{(u)},u}) C_v^{(0)} = E_u^{(0)}$, when $x_{i_{\text{nrg}}^{(u)},u} = 1$, since, by constraint (31), we have that $x_{v,u} = x_{i_{\text{nrg}}^{(u)},u}$, for any $v \leq_{\sigma_u} i_{\text{nrg}}^{(u)}$. On the other hand, when $x_{i_{\text{nrg}}^{(u)},u} = 0$, charger u will not be able to spend all of its energy, since the nodes it can reach cannot store all of it. This is also captured by our objective function, since $E_u^{(0)} x_{i_{\text{nrg}}^{(u)},u} + \sum_{v \leq_{\sigma_u} i_{\text{nrg}}^{(u)}} (x_{v,u} - x_{i_{\text{nrg}}^{(u)},u}) C_v^{(0)} = \sum_{v \leq_{\sigma_u} i_{\text{nrg}}^{(u)}} x_{v,u} C_v^{(0)}$, when $x_{i_{\text{nrg}}^{(u)},u} = 0$, which is equal to the total energy that the nodes reachable from u could harvest in total.

In our experimental evaluation, we solve IP-LRDC by first making a linear relaxation and then rounding the solution so that the constraints (30), (31) and (32). It is easy to see that the objective function value that we get is a lower bound on the optimal solution of the LREC problem. We use this bound to evaluate the performance of our iterative algorithm IterativeLREC (see Sect. 5.3).



(a) ChargingOriented (b) IterativeLREC (c) IP-LRDC

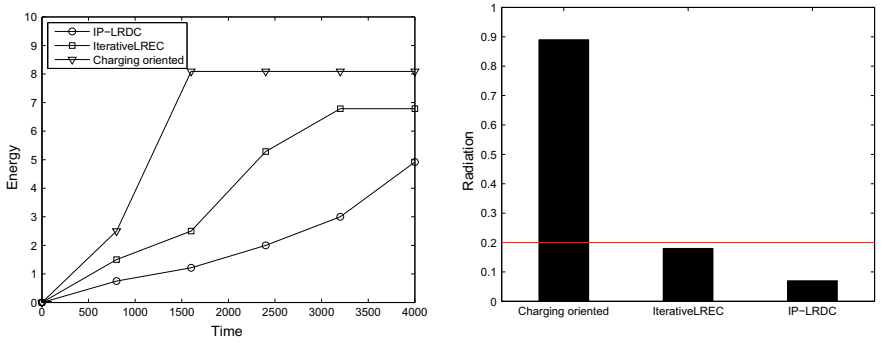
Fig. 9 Network snapshot using five chargers

5.5 *Performance Evaluation*

We conducted simulations in order to evaluate our methods using Matlab R2014b. We compared IterativeLREC, IP-LRDC (after the linear relaxation) and a charger configuration scheme in which each charger u sets its radius equal to $\text{dist}(u, i_{rad}^{(u)})$. We call this new configuration "ChargingOriented" because it assigns the maximum radius to each charger, without individually violating the radiation threshold. In other words, this configuration provides the best possible rate of transferring energy in the network and serves as an upper bound on the charging efficiency of the performance of IterativeLREC, but is expected to achieve a poor performance on keeping the radiation low, due to frequent, large overlaps. A snapshot of a uniform network deployment with $|\mathscr{P}| = 100$, $|\mathscr{M}| = 5$ and $K = 100$, is shown in Fig. 9. We observe that the radii of the chargers in the ChargingOriented case are larger than the other two cases. In the case of IP-LRDC, the radiation constraints lead to a configuration where two chargers are not operational. IterativeLREC provides a configuration in between the ChargingOriented and IP-LRDC, in which some overlaps of smaller size are present.

We deploy uniformly at random $|\mathscr{P}| = 100$ network nodes of identical capacity, $|\mathscr{M}| = 10$ wireless chargers of identical energy supplies and $K = 1000$ points of radiation computation. We set $\alpha = 0$, $\beta = 1$, $\gamma = 0.1$ and $\rho = 0.2$. For statistical smoothness, we apply the deployment of nodes in the network and repeat each experiment 100 times. The statistical analysis of the findings (the median, lower, and upper quartiles, outliers of the samples) demonstrate very high concentration around the mean, so in the following figures we only depict average values. We focus our simulations on three basic metrics: charging efficiency, maximum radiation and energy balance.

Charging efficiency. The objective value that is achieved as well as the time that is spent for the charging procedure is of great importance to us. The objective values achieved were 80.91 by the ChargingOriented, 67.86 by the IterativeLREC and 49.18 by the IP-LRDC. The ChargingOriented method is the most efficient and quick, as expected but it results in high maximum radiation. As we observe in Fig. 10a, it distributed the energy in the network in a very short



(a) Charging efficiency over time.

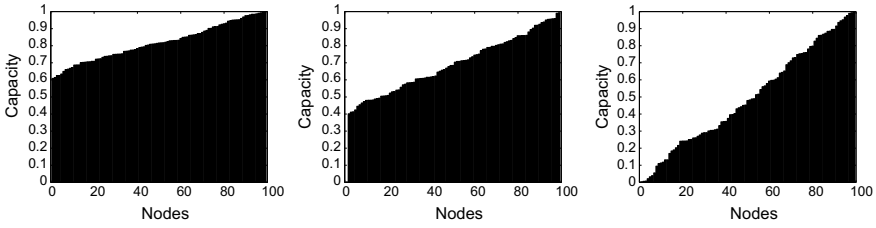
(b) Maximum radiation.

Fig. 10 Efficiency and radiation

time. The efficiency of `ChargingOriented` both in terms of objective value and in terms of time is explained by the frequent charger radii overlaps that are created during the configuration of the chargers (e.g., Fig. 9). `IP-LRDC` achieves the lowest efficiency of all due to the small charging radii and consequently small network coverage by the chargers. Our heuristic `IterativeLREC` achieves high enough efficiency w.r.t. the radiation constraints. Its performance lies between the performance of `ChargingOriented` and `IP-LRDC`, both in terms of objective value and in terms of time.

Maximum radiation. The maximum amount of radiation incurred is very important regarding the safety impact of the corresponding charging method. High amounts of radiation, concentrated in network regions may render a method nonpractical for realistic applications. This is the case for the `ChargingOriented`, which in spite of being very (charging) efficient, it significantly violates the radiation threshold (Fig. 10b). `IterativeLREC` is performing very well, since it does not violate the threshold but in the same time provides the network with high amount of energy.

Energy balance. The energy balance property is crucial for the lifetime of Wireless Distributed Systems, since early disconnections are avoided and nodes tend to save energy and keep the network functional for as long as possible. For this reason, apart from achieving high charging efficiency, an alternative goal of a charging method is the balanced energy distribution among the network nodes. Figure 11 is a graphical depiction of the energy provided in the network throughout the experiment. The nodes are sorted by their final energy level and by observing the figure, we are able to make conclusions about the objective value and the energy balance of each method. Our `IterativeLREC` achieves efficient energy balance that approximates the performance of the powerful `ChargingOriented`.



(a) `ChargingOriented` (b) `IterativeLREC` (c) `IP-LRDC`

Fig. 11 Energy balance

6 Conclusions

In this chapter, we presented two quite distinct approaches for radiation control in wireless distributed systems. In particular, we first studied the minimum radiation path problem of finding the lowest radiation trajectory of a person moving from a source to a destination point within the area of a network of wireless devices. Second, we studied the problem of efficiently charging a set of rechargeable nodes using a set of wireless energy chargers, under safety constraints on the electromagnetic radiation incurred. For both these problems, we provided hardness indications and theoretical results providing helpful insights. Furthermore, we presented and analyzed efficient algorithms and heuristics for approximating optimal solutions, namely minimum radiation trajectories and charging schemes respectively. Finally, we provided and discussed our experimental findings that verify our analysis and theoretical results.

References

1. Angelopoulos, C.M., Nikoletseas, S., Patroumpa, D., Raptopoulos, C.: Radiation-aware data propagation in wireless sensor networks. In: Proceedings of the 10th ACM International Symposium on Mobility Management and Wireless Access, MobiWac '12, pp. 11–18 (2012)
2. Angelopoulos, C.M., Nikoletseas, S., Raptis, T.P.: Wireless energy transfer in sensor networks with adaptive, limited knowledge protocols. Comput. Netw. **70**, 113–141 (2014)
3. Angelopoulos, C.M., Nikoletseas, S., Raptis, T.P., Raptopoulos, C., Vasilakis, F.: Improving sensor network performance with wireless energy transfer. Int. J. Ad Hoc Ubiquitous Comput. **20**(3), 159–171 (2015)
4. Angelopoulos, C.M., Nikoletseas, S.E., Patroumpa, D., Raptopoulos, C.: Radiation-aware data propagation in wireless sensor networks. In: Proceedings of the 10th ACM International Symposium on Mobility Management and Wireless Access (MOBIWAC), pp. 11–18 (2012)
5. Clementi, A.E.F., Pasquale, F., Silvestri, R.: MANETS: high mobility can make up for low transmission power. In: Proceedings of the 36th International Colloquium on Automata, Languages and Programming (ICALP), pp. 387–398 (2009)
6. Dai, H., Liu, Y., Chen, G., Wu, X., He, T.: Safe charging for wireless power transfer. In: INFOCOM, 2014 Proceedings IEEE, pp. 1105–1113 (2014). https://doi.org/10.1109/INFOCOM.2014.6848041

7. Dai, H., Liu, Y., Chen, G., Wu, X., He, T.: Safe charging for wireless power transfer. In: Proceedings of the IEEE Conference on Computer Communications (INFOCOM), pp. 1105–1113 (2014)
8. Dai, H., Liu, Y., Chen, G., Wu, X., He, T.: SCAPE: safe charging with adjustable power. In: 2014 IEEE 34th International Conference on Distributed Computing Systems (ICDCS), pp. 439–448 (2014). https://doi.org/10.1109/ICDCS.2014.52
9. Dolev, S., Gilbert, S., Guerraoui, R., Kuhn, F., Newport, C.C.: The wireless synchronization problem. In: Proceedings of the 28th Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC), pp. 190–199 (2009)
10. Edwards, M.J., Saunders, R.D., Shiota, K.: Effects of heat on embryos and foetuses. Int. J. Hyperth. **19**(3), 295–324 (2002)
11. Erlebach, T., Grant, T.: Scheduling multicast transmissions under SINR constraints. In: Proceedings of the 6th International Workshop on Algorithms for Sensor Systems, Wireless Ad Hoc Networks and Autonomous Mobile (ALGOSENSORS), pp. 47–61 (2010)
12. Erlebach, T., van Leeuwen, E.J.: Approximating geometric coverage problems. In: Proceedings of the ACM-SIAM Symposium on Discrete Algorithms (SODA), pp. 1267–1276 (2008)
13. Fei, X., Boukerche, A., de Araujo, R.B.: Irregular sensing range detection model for coverage based protocols in wireless sensor networks. In: Proceedings of the Global Communications Conference (GLOBECOM), pp. 1–6 (2009)
14. Gandhi, O., Morgan, L., de Salles, A., Han, Y., Herberman, R., Davis, D.: Exposure limits: the underestimation of absorbed cell phone radiation, especially in children. Electromagn. Biol. Med. **31**(1), 34–51 (2012)
15. Garey, M., Johnson, D., Stockmeyer, L.: Some simplified NP-complete graph problems. Theor. Comput. Sci. **1**(3), 237–267 (1976)
16. Guo, S., Wang, C., Yang, Y.: Joint mobile data gathering and energy provisioning in wireless rechargeable sensor networks. IEEE Trans. Mob. Comput. **13**(12), 2836–2852 (2014). https://doi.org/10.1109/TMC.2014.2307332
17. Han, S.Y., Abu-Ghazaleh, N.B.: A realistic model of co-located interference for wireless network packet simulation. In: Proceedings of the 7th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (MASS), pp. 472–481 (2010)
18. Havas, M., Marrongelle, J., Pollner, B., Kelley, E., Rees, C., Tully, L.: Provocation study using heart rate variability shows microwave radiation from 2.4 GHz cordless phone affects autonomic nervous system. Eur. J. Oncol. Libr. **5** (2010)
19. Intel: wireless resonant energy link (WREL) demo. http://software.intel.com/en-us/videos/wireless-resonant-energy-link-wrel-demo/
20. Jarry, A., Leone, P., Nikoletseas, S.E., Rolim, J.D.P.: Optimal data gathering paths and energy balance mechanisms in wireless networks. In: Proceeding of the 6th IEEE International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 288–305 (2010)
21. Jiang, L., Wu, X., Chen, G., Li, Y.: Effective on-demand mobile charger scheduling for maximizing coverage in wireless rechargeable sensor networks. Mob. Netw. Appl. **19**(4), 543–551 (2014)
22. Johnston, A.: Reliability and Radiation Effects in Compound Semiconductors. World Scientific Publishing Co. Pte. Ltd. (2010)
23. Kaklamanis, C., Kirousis, L.M., Bose, P., Kranakis, E., Krizanc, D., Peleg, D.: Station layouts in the presence of location constraints. In: Algorithms and Computation. Lecture Notes in Computer Science, vol. 1741, pp. 269–278. Springer, Berlin, Heidelberg (1999)
24. Kesselheim, T., Vocking, B.: Distributed contention resolution in wireless networks. In: Proceedings of the 24th International Symposium on Distributed Computing (DISC), pp. 163–178 (2010)
25. Kowalski, D.R., Pelc, A.: Optimal deterministic broadcasting in known topology radio networks. Distrib. Comput. **19**(3), 185–195 (2007)
26. Kurs, A., Karalis, A., Moffatt, R., Joannopoulos, J.D., Fisher, P., Soljacic, M.: Wireless power transfer via strongly coupled magnetic resonances. Science **317**, 83 (2007)

27. Leone, P., Nikoletseas, S.E., Rolim, J.D.P.: An adaptive blind algorithm for energy balanced data propagation in wireless sensors networks. In: Proceedings of the International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 35–48 (2005)
28. Li, Z., Peng, Y., Zhang, W., Qiao, D.: J-RoC: a joint routing and charging scheme to prolong sensor network lifetime. In: Proceedings of the 2011 19th IEEE International Conference on Network Protocols, ICNP '11, pp. 373–382 (2011)
29. Madhja, A., Nikoletseas, S., Raptis, T.P.: Distributed wireless power transfer in sensor networks with multiple mobile chargers. Comput. Netw. **80**, 89–108 (2015)
30. Madhja, A., Nikoletseas, S., Raptis, T.P.: Hierarchical, collaborative wireless charging in sensor networks. In: Proceedings of the IEEE Wireless Communications and Networking conference, WCNC '15 (2015)
31. Martirosyan, A., Pazzi, A.B.R.W.N.: Energy-aware and quality of service-based routing in wireless sensor networks and vehicular ad hoc networks. Ann. des Telecommun. **63**(11–12), 669–681 (2008)
32. Murata: Murata manufacturing. http://www.murata.com/
33. Nikoletseas, S., Patroumpa, D., Prasanna, V., Raptopoulos, C., Rolim, J.: Radiation awareness in three-dimensional wireless sensor networks. In: 2012 IEEE 8th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 176–185 (2012)
34. Nikoletseas, S., Raptis, T.P., Souroulagkas, A., Tsolovos, D.: An experimental evaluation of wireless power transfer protocols in mobile ad hoc networks. In: Proceedings of the IEEE Wireless Power Transfer Conference, WPTC '15 (2015)
35. Nikoletseas, S.E., Patroumpa, D., Prasanna, V.K., Raptopoulos, C., Rolim, J.D.P.: Radiation awareness in three-dimensional wireless sensor networks. In: Proceedings of the IEEE 8th International Conference on Distributed Computing in Sensor Systems (DCOSS), pp. 176–185 (2012)
36. Nikoletseas, S.E., Raptis, T.P., Raptopoulos, C.: Low radiation efficient wireless energy transfer in wireless distributed systems. In: Proceedings of the 35th IEEE International Conference on Distributed Computing Systems (ICDCS), pp. 196–204 (2015)
37. Ntzouni, M., Skouroliakou, A., Kostomitsopoulos, N., Margaritis, L.: Transient and cumulative memory impairments induced by GSM 1.8 GHz cell phone signal in a mouse model. Electromagn. Biol. Med. **32**(1), 95–120 (2013)
38. Oltenau, M., Marincas, C., Rafiroiu, D.: Dangerous temperature increase from em radiation around metallic implants. Acta Electroteh. **53**(2), 175–180 (2012)
39. Pettarin, A., Pietracaprina, A., Pucci, G., Upfal, E.: Tight bounds on information dissemination in sparse mobile networks. In: Proceedings of the 30th Annual ACM SIGACT-SIGOPS Symposium on Principles of Distributed Computing (PODC), pp. 355–362 (2011)
40. Powercast: Powercast corporation. http://www.powercastco.com/
41. TI: Texas instruments. http://www.ti.com/
42. WPC: The wireless power consortium. http://www.wirelesspowerconsortium.com/
43. Xiang, L., Luo, J., Han, K., Shi, G.: Fueling wireless networks perpetually: a case of multi-hop wireless power distribution. In: 2013 IEEE 24th International Symposium on Personal Indoor and Mobile Radio Communications (PIMRC), pp. 1994–1999 (2013). https://doi.org/10.1109/PIMRC.2013.6666471
44. Yu, Y., Prasanna, V.K.: Energy-balanced task allocation for collaborative processing in wireless sensor networks. Mob. Netw. Appl. **10**(1–2), 115–131 (2005)

Subspace-Based Encryption

Atef Mermoul and Adel Belouchrani

Abstract The use of signal processing in the field of cryptography is an attractive approach in the recent years. As an example, the concept of Blind Source Separation (BSS) has been used for speech encryption. However, some weaknesses of this proposal from a cryptographic point of view have been observed. This chapter proposes to bypass the weaknesses, from a security point of view, of such encryption schemes based on blind source separation techniques by using vectorial subspace concepts. This leads to the proposal of new subspace-based encryption systems. The new subspace-based encryption method together with its iterative version are designed and analyzed in terms of security robustness and quality of recovered signals. Security analysis is conducted using cryptanalysis attacks, whereas quality assessment is achieved using both objective and subjective tools. The proposed subspace-based encryption system is applied for speech and images. Experimental results show an enhancement in the performances beside some interesting and specific features, from a cryptographic point of view, brought by the proposed encryption system.

1 Introduction

The rapid growth of digital communications and electronic data exchange make information security a crucial issue in industry, business, and administration. Modern cryptography provides essential techniques for securing information and protecting data [1]. Among cryptographic techniques, Blind Source Separation (BSS)-based methods have received some attention in speech and image encryption fields [2–9]. The latter exploits the concept of blind source separation, which consists of recovering mutually independent source signals from their mixtures without any knowledge

A. Mermoul · A. Belouchrani (✉)
Electrical Engineering Department/LDCCP, Ecole Nationale Polytechnique, Algiers, Algeria
e-mail: adel.belouchrani@enp.edu.dz

A. Mermoul
e-mail: atef.mermoul@enp.edu.dz

H. M. Ammari (ed.), *Mission-Oriented Sensor Networks and Systems: Art and Science*, Studies in Systems, Decision and Control 164,
https://doi.org/10.1007/978-3-319-92384-0_21

of the mixing coefficients [10, 11]. Several encryption techniques based on such BSS concept have been introduced [2–4]. In image encryption, the transmitted images are hidden in a noise image by specific mixing before encryption and then recovered through BSS after decryption [2]. In speech encryption, a time domain scrambling scheme is used to mask the speech signal with a random noise by specific mixing [4]. Another speech encryption technique takes advantage of the underdetermined BSS problem [5].

When looking in details to such techniques, it appears that BSS techniques are more suitable for cryptanalysis purposes rather than for cryptographic ones. This is due essentially to the fact that BSS techniques are, by their definition, tools developed to recover unknown signals from their observed mixtures without the knowledge of the mixing coefficients. This is, by analogy, the same formulation of the cryptanalysis problem, i.e., recovering a plain-text (or a set of plain-texts) from cipher-texts (mixtures of plain-texts and cryptographic keys) without knowing the cryptographic keys (mixing coefficients). In [9], the security weaknesses and defects of BSS-based encryption schemes are established from a cryptographic point of view. These observations on safely using BSS-based techniques in cryptographic field gave a stimulus to develop new techniques, which would bypass their actual security drawbacks.

To overcome the limitations of the BSS-based encryption schemes revealed in [9], we have introduced in [10–12] new speech encryption techniques based on the subspace concept. Details on the latter together with the image encryption application are the core added value of this book chapter. An assessment methodology is applied for the new subspace-based encryption technique to appreciate its quality and security levels. Several tests and evaluations are conducted to assess the cryptographic robustness of this class of techniques. Hence, this chapter focuses on studying and analyzing the use of the subspace concept by investigating first the opportunity of using blind source separation techniques in the encryption domain. We will be discussing the various constraints related to the performance of these techniques. The main tasks are to survey the current status, identify the limitations of these techniques, and propose alternative approaches. The analysis would approach the various aspects of the security of blind source techniques used in the encryption domain and their performance. We are looking forward to provide a new research direction towards subspace-based techniques to bypass the limitations and drawbacks inherent to the BSS techniques used in encryption field. Specifically, we would be focusing on the security aspects of such techniques.

This chapter is organized as follows: Sect. 2 introduces a brief background on cryptographical and cryptanalysis techniques besides a general reminder on the key concepts of linear algebra used throughout the chapter. A state of the art of the use of blind source separation techniques in encryption field with their cryptanalysis results are presented in Sect. 3. Starting from the cryptographic weaknesses of blind source separation approach, Sect. 4 introduces the subspace-based encryption techniques. The proposed encryption system based on subspace concept is studied in detail. An iterative version of the subspace scheme is presented in Sect. 5. In Sect. 6, several tests using cryptanalysis attacks are conducted on the subspace-based encryption systems to evaluate their robustness from a security point of view. Section 7 presents

the results of the experiments conducted on the subspace-based encryption method and its iterative version with application to speech and image encryptions. The performance is evaluated in term of quality and security. Section 8 concludes the chapter.

2 Preliminaries

2.1 Cryptography

Hiding some information or making it incomprehensive to others is a very old human need. Several means were used to meet this need but the process of putting the bases of a whole science, nowadays called as cryptology, started only in the seventh century. If cryptology experienced all this development several centuries before, it is because it met partly quite precise needs of the society/state of that time and even anticipated the future needs in precise fields [13, 14]. Cryptography has been a restricted area controlled only by military and diplomatic entities throughout the world. That is why it had and still has, somewhat, a specific reputation. However, during the past decades, the fast development of information and communications technologies causes a widespread use of cryptology tools. It has been implemented in various equipment and devices, by software and hardware means.

Usually, in the cryptography literature, the term plain-text is used to refer to the message to be transmitted over communications medium, whatever its nature is. It could be a text, audio, video, or data. After encryption, it becomes a cipher-text. A general descriptive equation of an encryption operation is given as

$$x = E(k_e, p), \tag{1}$$

where x is the cipher-text, p is the plain-text, k_e is the key (encryption parameter), and E is the encryption algorithm. On the receiving side, to recover the plain-text, the cipher-text c is decrypted using a decryption algorithm D as follows:

$$p = D(k_d, x), \tag{2}$$

where k_d is the decryption key. k_e and k_d could be either different or the same. It depends on the type of the cryptography system.

2.1.1 Classical Cryptography

The objective of classical cryptography is to guarantee the confidentiality of the plain-text to be encrypted and sent to a receiver. The principles of perfect secrecy as shown by C. Shannon, in 1949, in his mathematical treatment entitled “Communication Theory of Secrecy Systems”, require that the encryption key length must be at least

the same as the plain-text length [15]. The encryption key has also to be randomly generated and used once. This ensures a perfect secrecy or what it is called "one-time pad" or Vernam cipher. In classical cryptography, the encryption scheme has always been a symmetric one in the sense that both sender and receiver have to share the same encryption keys before starting exchanging encrypted messages. This requirement procures a high degree of confidentiality in the case, where the encryption keys have been "correctly" generated and distributed to both sender and receiver. However, from an operational point of view, the management of such a scheme becomes very hard in the presence of an important number of users, senders, and receivers, who have to exchange encrypted messages.

2.1.2 Modern Cryptography

Modern cryptography is based on complexity theory mainly the "computational complexity". It assumes true randomness together with one-way functions. The inversion of the aforementioned functions is computationally intractable. The true randomness can be well approximated by pseudorandomness, i.e., the one provided by computers. Modern cryptography schemes can possibly be asymmetric, i.e., using different keys for encryption and decryption. One of the first asymmetric schemes was reported in [16] and its first implementation appeared in [17].

2.2 Cryptanalysis

Cryptanalysis is the second half of cryptology; science which includes cryptography. The desire of knowing the secrets of other persons or groups, which use cryptographic tools to secure their communications gives arise to cryptanalysis. For a long time, the confrontation between cryptography and cryptanalysis was occurring on a pure mathematical ground. Mathematical solutions for securing correspondences were defeated by other mathematical tools [18]. In the earlier cryptographic techniques such as alphabetical substitutions or permutations, cryptanalysis was based on frequency analysis of the used languages. Except brute force attack which remains the last approach to use because of its time and computing resource consumption, some recent techniques have proven to be very efficient against several cryptographic algorithms. As the cryptographic algorithms become more complex, the cryptanalysis becomes more difficult. To reduce this difficulty, new approaches have taken place [18]. Successful attacks may, for example, recover the plain-text (or parts of the plain-text) from the cipher-text, substitute parts of the original message, or forge the digital signatures [1]. Nowadays, providing evidence that the robustness of a cryptographic algorithm is not as it was claimed is a successful attack even though it does not recover, fully or partially, any of the plain-text or the encryption key.

2.2.1 A Brief Historical View

It would be interesting to have a brief view on the history of cryptanalysis. During the campaign of translation of books and manuscripts written in several difficult and old languages, and sometimes in dead languages, there was a pressing need to master all known cryptographic tools and techniques. Some of these books and manuscripts, especially in certain areas like chemistry and magic, contained encrypted paragraphs. This need gave rise to a new science: cryptanalysis [13]. A research group (using modern terminology) under the supervision of Yakoob Ibn Ishak Al-Kindi, known as Alkindus, worked at Bait Al-Hikmah in Baghdad, on decrypting the encrypted paragraphs in order to complete the translation process of all the submitted manuscripts [13]. They were the first to discover and write down the methods of cryptanalysis [19]. Among the 290 manuscripts he wrote in various fields, appears the oldest one which discovered and wrote down the methods of cryptanalysis: Rissalatoon fi istikhradji al mooamma (a writing in extracting the encrypted) [13]. Al-Kindi founded the principles of cryptanalysis. He proposed four methods of decryption: quantitative techniques, qualitative techniques, probable word, and letters combination. In his manuscript Kitaboo al-moo amma (book of the encrypted), an important handbook of cryptology even centuries later, Al-Kindi proposed a classification diagram of encryption methods and their related cryptanalysis techniques [13]. On another hand, other conditions supported the emancipation of this new science. Disciplines that were developed at that time, like grammar and mathematics, had considerable contribution. Cryptanalysis had an enormous requirement for tools of analyzing languages in which the encrypted texts were written. This helped the mastering of the qualitative approach in cryptanalysis. As for the quantitative approach, like calculating letter frequencies of several languages, mathematics was very well developed. Centuries later, the Second World War balanced because of a cryptanalysis team's hard work at Bletchley Park in the UK. They broke Enigma, the famous German encryption machine and got the ability to "read" the confidential messages exchanged within German army. They got the possibility to know, e.g., the plans and the positions.

2.3 Cryptanalysis Attacks

Below, a general classification of cryptanalysis attacks is provided [1].

2.3.1 Cipher-Text-Only Attack

This is the most general attack where the attacker has access only to cipher-text. Since cipher-texts are sent and received via communications mediums (e.g., networks, radio, satellites), one has to suppose the availability, by default, of all cipher-texts to potential attackers. So, this attack should be considered for every cryptographic

algorithm assessment and is considered as the basic level for security robustness evaluation.

2.3.2 Known-Plain-Text Attack

In this type of attack, it is assumed that the attacker can get pairs of plain-text–cipher-text. The attack consists of trying to decrypt the cipher-text using information extracted and gathered from pairs of plain-text–cipher-text. Using the information extracted from these pairs, the attacker attempts to decrypt a cipher-text for which he does not have the plain-text. The use of standard formats of messages could be useful to the attacker in conducting known-plain-text attack [1].

2.3.3 Chosen-Plain-Text Attack

In this type of attack, it is assumed that the attacker can encrypt plain-texts of his choice and get their corresponding cipher-texts. Naturally, to realize such an attack, the attacker has to get access, at least, once to the encryption device [1]. Then, the cryptanalysis work consists of trying to decrypt cipher-texts for which he does not have the corresponding plain-text.

2.3.4 Adaptively Chosen-Plain-Text Attack

This type of attack is similar to the chosen-plain-text attack except that here, the attacker can get more pairs of plain-text–cipher-text by doing some analysis and can have access as long as he wants to the encryption device [1].

2.3.5 Chosen-and Adaptively Chosen-Cipher-Text Attack

In this type of attack, the attacker has the ability to choose cipher-texts and then decrypt them to get the corresponding plain-texts. He needs to have access to the decryption device [1]. Despite the type of the cryptanalysis attack, a basic principle of cryptanalysis is to assume that the algorithm is not secret, i.e., it is well known by the attacker.

2.3.6 Non-classical Cryptanalysis Approaches

The robustness of a cryptographic application depends not only on its pure mathematical model, but also on its implementation on soft and/or hardware devices. Some parameters which are not involved in the mathematical aspect of cryptographic solutions and, hence, tend to be ignored in the security evaluation such as execution time

and power consumption can be very important and reveal secret information. This can cause the break of a, theoretically secure, cryptographic algorithm [18]. "Side-channel attacks" are attacks that exploit this side-channel information to retrieve the secret information treated by cryptographic devices. Several types of side-channel attacks are published in the literature. They include timing attacks [20], power analysis attacks [21], electromagnetic attacks [22], fault induction attacks, and template attacks [23, 24]. Actually, Side-Channel Attacks exploit information that leaks from a cryptographic device [21].

2.4 A Brief Reminder on Some Basics on Linear Algebra

This is not a linear algebra section, however, it turns out that many important mathematical properties of cryptography and cryptanalysis are based on algebraic concepts [25]. That is why a brief reminder is necessary for eliminating any reader confusion on the cryptographic construction explained in this chapter.

2.4.1 Subspace

The space $\mathscr{H}$ spanned by a collection of vectors $\{\mathbf{x}_k\}$

$$\mathscr{H} := \{\alpha_1\mathbf{x}_1 + \cdots + \alpha_n\mathbf{x}_n | \alpha_i \in \mathbf{C}, \forall i\} \tag{3}$$

is called a *linear subspace*.

2.4.2 Basis

An independent collection of vectors that together span a subspace is called a *basis* for that subspace. If the vectors are orthogonal (${\mathbf{x}_i}^H\mathbf{x}_j = 0, i = j$), it is an *orthogonal basis*.

2.4.3 Projection

A square matrix $\mathbf{P}$ is a projection if $\mathbf{PP} = \mathbf{P}$. It is an orthogonal projection if also $\mathbf{P}^H = \mathbf{P}$.

- The norm of an orthogonal projection is $\|\mathbf{P}\| = 1$.
- For an isometry $\hat{\mathbf{U}}$, the matrix $\mathbf{P} = \hat{\mathbf{U}}\hat{\mathbf{U}}^H$ is an orthogonal projection (onto the space spanned by the columns of $\hat{\mathbf{U}}$).

2.4.4 Notations

T, H, $\sharp$ denote transpose, conjugate-transpose, Moore–Penrose pseudoinverse, respectively. Matrices and vectors denoted with boldface type, using capital letters for matrices and lower case letters for vectors. Given a matrix $\mathbf{A} \in \mathbb{C}^{N\times r}$, the range subspace of $\mathbb{C}^N$ spanned by the $r \leq N$ columns of $\mathbf{A}$ is denoted by $\langle A \rangle$.

3 Blind Source Separation in Cryptography

In this section, we present an overview of the use of blind source separation techniques in the cryptography field.

3.1 Blind Source Separation (BSS)

Blind Source Separation (BSS) aims to recover a set of unknown mutually independent source signals from their observed mixtures without any knowledge of the mixing coefficients [10, 11]. Suppose that there exists M independent source signals and N observed mixtures of the source signals (usually $M \leq N$). The linear mixing model is as follows:

$$\mathbf{x}(t) = \mathbf{H}\mathbf{s}(t), \tag{4}$$

where $\mathbf{s}(t) = [s_1(t), \ldots, s_M(t)]^T$, which is an $M \times 1$ column vector collecting the source signals, vector $\mathbf{x}(t)$ similarly collects the N observed (mixed) signals, and $\mathbf{H}$ is an $N \times M$ mixing matrix that contains the mixing coefficients. The purpose of BSS is to find an $M \times N$ unmixing matrix $\mathbf{W}$ such that the $M \times 1$ output vector $\mathbf{u}(t)$ verifies

$$\mathbf{u}(t) = \mathbf{W}\mathbf{x}(t) = \mathbf{W}\mathbf{H}\mathbf{s}(t) = \mathbf{P}\mathbf{D}\mathbf{s}(t), \tag{5}$$

where $\mathbf{P}$ and $\mathbf{D}$ denote a permutation matrix and diagonal matrix, respectively. When $M \leq N$, the blind source separation and more specifically source recovery is possible. However, when $M > N$, BSS becomes generally impossible except under specific conditions [26–28]. This is referred to as the underdetermined BSS problem.

3.2 BSS-Based Encryption

Blind Source Separation (BSS)-based techniques have been used in speech and image encryption fields [2–4, 29]. In Refs. [2] and [3], the BSS mixture model is exploited in image encryption. Before encryption, the original images are hidden in a noise image through specific mixture coefficients. The original images are then recovered

by BSS after decryption. Reference [4] introduces a speech encryption algorithm that integrates a time domain scrambling scheme in order to mask the speech signal with a random noise using specific mixture coefficients. In Ref. [5], a speech encryption algorithm is proposed based on the underdetermined BSS problem. The encryption procedure of this algorithm goes as follows:

$$\mathbf{x}(t) = \mathbf{A}_p \mathbf{p}(t) + \mathbf{A}_k \mathbf{k}(t) \tag{6}$$

where $\mathbf{p}(t) = [p_1(t), \ldots, p_M(t)]^T$ and $\mathbf{k}(t) = [k_1(t), \ldots, k_M(t)]^T$ represent M input plain-signals and M key signals, respectively. $\mathbf{A}_p$ and $\mathbf{A}_k$ are $M \times M$ matrices, both of which elements are within $[-1, 1]$. The decryption procedure, as long as $\mathbf{A}_s$ is invertible, is given by

$$\mathbf{p}(t) = \mathbf{A}_p^{-1} \left(\mathbf{x}(t) - \mathbf{A}_k \mathbf{k}(t)\right). \tag{7}$$

In the BSS-based encryption scheme [5], the key signals $k_1(t), \ldots, k_M(t)$ are as long as the plain-signals and have to be generated by a Pseudorandom Number Generator (PRNG) with a secret seed, which serves as the secret key. The mixing matrices $\mathbf{A}_s$ and $\mathbf{A}_k$, being secret parameters, may be known by the receiver as secret keys and hence their estimation by a BSS approach at the receiver should not be necessary. Hence, the BSS approach is, in this case, worth to be used in a cryptanalysis process rather than in an encryption one. However, some weaknesses from a cryptographic point of view exist and the security against some attacks is not sufficiently strong [9]. The encryption procedure described in Eq. (6) could be presented under the form of two steps as follows:

- *Step 1*: $\mathbf{x}^{(1)}(t) = \mathbf{A}_p\ \mathbf{p}(t)$;
- *Step 2*: $\mathbf{x}(t) = \mathbf{x}^{(1)}(t) + \mathbf{A}_k \mathbf{k}(t)$.

As it is presented above, one can see that this procedure is equivalent to a simple matrix-based block cipher in the first step and a simple addition-based stream cipher. The security of this BSS-based encryption scheme is analyzed in the following section.

3.3 Cryptanalysis of BSS-Based Encryption

Herein, the security weaknesses and defects of BSS-based encryption scheme are discussed more precisely in term of its weaknesses against known/chosen-plain-text attack and chosen-cipher-text attack [9].

3.3.1 The Mixing Matrix A

As long as the principles of BSS techniques are respected, the mixing matrix $\mathbf{A}$ seems to be not required at the decryption side to separate the encrypted signals [9].

However, if it is so, i.e., $\mathbf{A}$ is not a secret parameter and considering that $\mathbf{x}^*(t) = \mathbf{A}_p^{-1}\mathbf{x}(t)$ is the equivalent obtained encrypted signal to the encryption procedure described by Eq. (6), the encryption procedure could hence be given by

$$\mathbf{x}^*(t) = \mathbf{p}(t) + \mathbf{A}_p^{-1}\mathbf{A}_k(t) \tag{8}$$

As it is shown in the the encryption procedure given by Eq. (8), there is no underdetermined BSS problem [9].

On another hand, if the mixing matrix $\mathbf{A}$ is not a secret parameter, the BSS-based encryption scheme would be in front of the problem of closely related input signals as it is the case in an image and its watermarked version. This difficulty is due to an essential hypothesis in BSS systems: the input signals are mutually independent of each other. Thus, it is clear that the mixing matrix $\mathbf{A}$ must be part of secret parameter used in BSS encryption scheme [9].

3.3.2 How Key Space is Large?

A mixing matrix $\mathbf{A}$ of dimension $P \times Q$, the secret key parameter of BSS-based encryption scheme, has its entries in the interval $[-1, 1]$ [2–8]. So, the number of all possible mixing matrix $\mathbf{A}$ is $R^{(P+Q)}$, where R is determined by the finite precision under which the cryptosystem is realized. P and Q are the number of input plain-signals and the number of key signals, respectively. Note that the BSS-based encryption scheme is mainly based on the principle of creating an underdetermined case by constructing a vector which contains both the plain-signals and the key signals. This gives rise to a $(P + Q)$ cipher-signals and leads to a $R^{(P+Q)}$ possible mixing matrices. For example, if the cryptosystem is implemented with n-bits fixed-point arithmetic, $R = 2^n$; if it is implemented with $IEEE$ floating-point arithmetic, $R = 2^{31}$ (single-precision) or $R = 2^{63}$ (double-precision) [9, 30]. Furthermore, the key signals $\mathbf{k}(t)$ are generated using a Pseudorandom Number Generator (PRNG) with a key seed $\mathbf{I}_0$, which has a length of J bits. This means that the size of the key space of key signals $\mathbf{k}(t)$ is 2^J. Thus, the size of the whole key space of the BSS-based encryption scheme is $R^{P(P+Q)}2^J$. In the case where $\mathbf{A} = [\mathbf{B}, \beta\mathbf{B}]$, the size of the key space is $R^{P^2}2^J$ [9].

3.3.3 Divide-and-Conquer (DAC) Attack

The encryption procedure described in Eq. (6) could be rewritten as

$$\mathbf{p}(t) \;=\; \hat{\mathbf{A}}\,\mathbf{x}_k(t) \tag{9}$$

where $\mathbf{x}_k(t) = [x_1(t), \ldots, x_P(t), k_1(t), \ldots, k_Q(t)]^T$ and

$$\hat{\mathbf{A}} = \mathbf{A}_p^{-1}\,[\mathbf{I}, -\mathbf{A}_k] \;=\; [\mathbf{A}_p^{-1}, -\mathbf{A}_p^{-1}\mathbf{A}_k] \tag{10}$$

As it can be seen from the above equation, the knowledge of $\mathbf{k}(t)$ and the ith row of $\hat{\mathbf{A}}$ allows recovering $p_i(t)$. This means that a Divide-and-Conquer attack (DAC) could separately break P rows of $\hat{\mathbf{A}}$. Hence, the number of possible mixing matrices becomes $PR^{(P+Q)}$ rather than $R^{P(P+Q)}$. Consequently, the size of the whole key space will be $PR^{(P+Q)}2^J$ rather than $R^{P(P+Q)}2^J$ [9].

3.3.4 Sensitivity to the Mixing Matrix A

A good cryptosystem should have a high sensitivity to key mismatch. This means that if two slightly different encryption keys are used to encrypt the same plain-text, the obtained cipher-texts should be very different [9, 31]. In the BSS encryption scheme, considering two mixing matrices $\mathbf{A}_1 = [a_{1;i,j}]$ and $\mathbf{A}_2 = [a_{2;i,j}]$ of size $M \times N$, if ε is the maximal difference of all elements, then Δx_i, the ith element of $\Delta\mathbf{x}$ is given by

$$\Delta\mathbf{x} = \mathbf{A}_1\mathbf{p}(t) - \mathbf{A}_2\mathbf{p}(t). \tag{11}$$

One can see that $\Delta\mathbf{x}$ verifies the following inequality:

$$\begin{aligned} |\Delta x_i| &= |\sum_{j=1}^{N}(a_{1;i,j} - a_{2;i,j})p_j| \\ &\leq \sum_{j=1}^{N}|a_{1;i,j} - a_{2;i,j}|.|p_j| \\ &\leq N\varepsilon\, max(|\mathbf{p}(t)|) \end{aligned} \tag{12}$$

where $|\mathbf{p}(t)|$ is the vector which contains absolute values of all elements of $\mathbf{p}(t)$, i.e., $|\mathbf{p}(t)| = [\,|p_1(t)|\ \ldots\ |p_N(t)|\,]^T$. The mixing matrix can be approximately guessed under a relatively large finite precision ε [9]. This low sensitivity of BSS-based encryption scheme to the mixing matrix is verified by the results obtained from encrypting a plain-text $\mathbf{p}(t)$ using a mixing matrix $\mathbf{A}$ and decrypting the resulting cipher-text $\mathbf{x}(t)$ using a mismatched mixing matrix $(\mathbf{A}, \varepsilon\mathbf{R})$, where $\varepsilon \in [0, 1]$ and $\mathbf{R}$ is a $P \times (P + Q)$. After decryption, one gets $\hat{\mathbf{p}}(t)$ which is an estimated version of $\mathbf{p}(t)$. The exhaustively search for an approximate version of the mixing matrix $\mathbf{A}$ under the finite precision $\varepsilon = 0.01$ allows to get a good estimation of the plain-texts [9].

Figure 1 shows a recovered plain-speech resulting from an exhaustively search of the mixing matrix $\mathbf{A}$ with a relatively large value of $\varepsilon = 0.1$. Besides the fact that experimental results confirm that a plain-text can be approximately recovered by a mismatched key, humans have a good capability of distinguishing images and speech even in presence of errors [9].

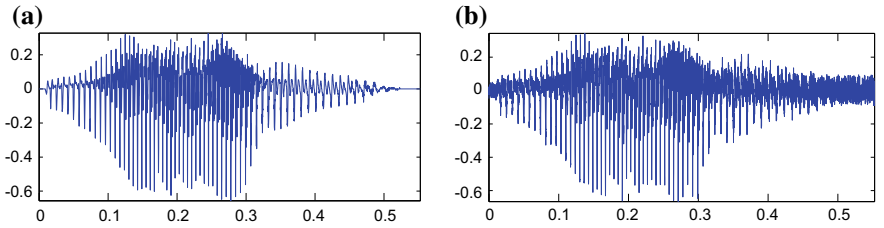


Fig. 1 A recovered speech from an exhaustively search of the mixing matrix $\mathbf{A}$ when $P = 2$ and $\mathbf{A} = [\mathbf{B}, \beta\mathbf{B}]$: **a** the original plain-speech, **b** the recovered speech

3.3.5 Sensitivity to the Key Signals $\mathbf{k}(t)$

The BSS-based encryption scheme has a low sensitivity to key signals mismatch because of the same reason of its low sensitivity to mixing matrix mismatch. If the maximal difference of all elements of two key signals $\mathbf{k}_1(t)$ and $\mathbf{k}_2(t)$ is ε, then each element of $|\mathbf{A}_k\mathbf{k}_1(t) - \mathbf{A}_k\mathbf{k}_2(t)|$ is not greater than $Q \max(|\mathbf{A}_k|)\varepsilon = Q\varepsilon$.

3.3.6 Differential Attack

The differential $\Delta_{\mathbf{x}}(t)$ between two cipher-texts $\mathbf{x}^{(1)}(t)$ and $\mathbf{x}^{(2)}(t)$ obtained from encrypting two plain-texts $\mathbf{p}^{(1)}(t)$ and $\mathbf{p}^{(2)}(t)$ using the same encryption key $(\mathbf{A}, \mathbf{I}_0)$ is given by

$$\Delta_{\mathbf{x}}(t) = \mathbf{A}_p^{-1}\Delta_{\mathbf{p}}(t) \tag{13}$$

where $\Delta_{\mathbf{x}}(t) = \mathbf{x}^{(1)}(t) - \mathbf{x}^{(2)}(t)$ and $\Delta_{\mathbf{p}}(t) = \mathbf{p}^{(1)}(t) - \mathbf{p}^{(2)}(t)$. Due to the low sensitivity of BSS-based encryption scheme to mixing matrix $\mathbf{A}$ as shown in Sect. 3.3, an exhaustive search could be applied to recover $\mathbf{A}_p$ [9]:

$$\Delta_{\mathbf{p}}(t) = \mathbf{A}_p^{-1}\Delta_{\mathbf{x}}(t) \tag{14}$$

In fact, the effect of key signals $\mathbf{k}(t)$ does not exist anymore when a differential attack is applied. Then, a mixed view of two interested plain-texts is obtained from the above calculation of the plain-text difference.

Low Sensitivity to Plain-text

In a good cryptosystem, the encryption of two plain-texts with a very slight difference should be very different [9]. However, in the BSS-based encryption scheme, when we use two very close plain-texts $\mathbf{p}_1(t)$ and $\mathbf{p}_2(t)$ for which the maximal difference of all elements is ε, then each element of $|\mathbf{A}_p\mathbf{p}_1(t) - \mathbf{A}_p\mathbf{p}_2(t)|$ is not greater than $P \max(|\mathbf{A}_p|)\varepsilon = P \times \varepsilon$. This low sensitivity increases when the two plain-texts are closely correlated as in the case of a plain-text and its watermarked version [9].

Known-plain-text Attack
By encrypting plain-texts with the same key, one can get in this type of attack plain-text differences. From Eq. (13), the mixing matrix can be determined using P plain-text differences as follows:

$$\mathbf{A}_p \;=\; \Delta_{\mathbf{x}}(t)(\Delta_{\mathbf{p}}(t))^{-1} \tag{15}$$

where $\Delta_{\mathbf{p}}(t)$ and $\Delta_{\mathbf{x}}(t)$ are $P \times P$ matrices, constructed row by row from the P plain-texts and the corresponding cipher-texts differences, respectively. Considering that n distinct plain-texts can generate $n(n-1)/2$ plain-text differences [9]. The number n of required plain-texts to yield at least P plain-text differences is given by $n \geq \sqrt{P}$ after solving the following inequality:

$$n \geq \lceil \sqrt{P - 1/4} \; + 1/2 \rceil \;\approx \sqrt{P} \tag{16}$$

Chosen-plain-text/cipher-text Attack
With a slight difference, the chosen-plain-text attack and the differential known-plain-text attack applied on BSS-based encryption scheme give roughly the same result [9]. In the chosen-cipher-text attack, one can choose a number of cipher-texts and observe the corresponding plain-texts.

3.4 Comments on Cryptanalysis of BSS-Based Encryption Scheme

The conducted cryptanalysis robustness study considers the cipher-text-only attack approach in terms of the resistance level to Divide-and-Conquer (DAC) attack and to differential attack, as well as the evaluation of sensitivity to the mixing matrix, to the key signals and to the plain-text. In known-plain-text attack approach, the number of required plain-texts to yield at least P plain-text differentials has been evaluated [9]. At this level, the analysis of the security robustness of BSS-based encryption scheme has shown that, in the actual architecture of this system, some weaknesses, from a cryptographic point of view, still exist. First, the key signals $\mathbf{k}(t)$ do not play any important security role in the case of a differential attack. This means that the effect of the second term of the encryption procedure described in Eq. (6) is canceled. Second, the use of the mixing matrix several times beside the low sensitivity of encryption/decryption constitute a weakness of BSS-based encryption scheme. However, from another point of view, in the BSS-based encryption scheme, the low sensitivity of decryption to cipher-text could be seen as an advantage in the case of the need to lossy decryption. Lossy decryption means that even when the receiver gets a cipher-text which is slightly different from the requested one, the decryption process could be achieved successfully. The lossy decryption is useful in some real applications, where the cipher-text could be compressed with some lossy

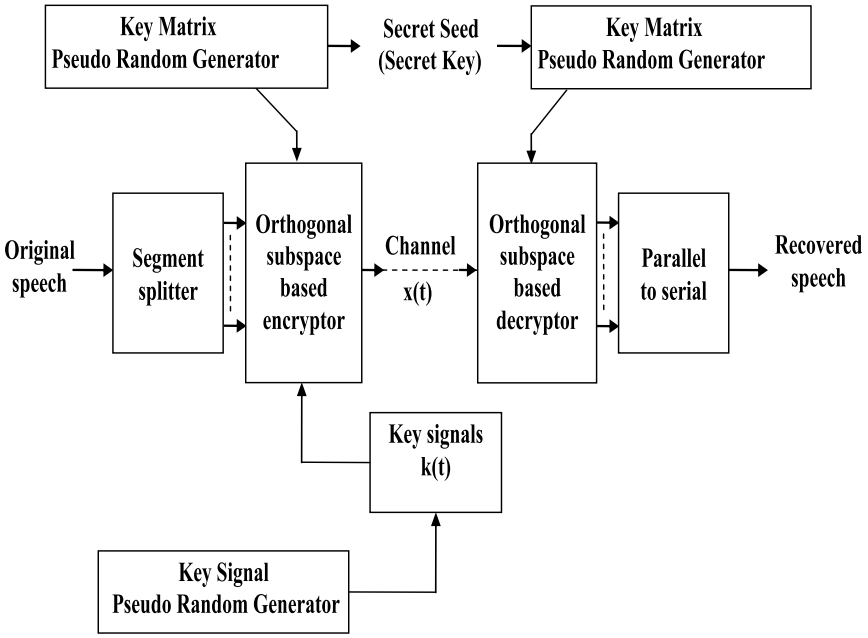


Fig. 2 Block diagram of the subspace-based encryption

algorithms to save the required storage. But, from a cryptographic point of view, this feature constitutes a considerable weakness [9].

Taking advantage of the security defects of BSS-based encryption scheme revealed by the cryptanalysis attacks, one can design a new encryption scheme based on subspace concept. The design of such a system mainly consists of the optimization of the following attributes:

- Cryptographic robustness based essentially on confusion and diffusion principles.
- Quality of restitution of the original signal after decryption.

4 Subspace-Based Encryption

The two main steps in an encryption scheme are the encryption and the decryption steps [32]. In practice, the output of the encryption step is transmitted through a communication channel and then received at the receiver's hand before being processed in the decryption step. Herein, the communication channel is assumed ideal and hence, the output of the encryption step is actually the input of the decryption step.

4.1 Encryption

The block diagram of the subspace-based encryption scheme is shown in Fig. 2. The data are first fed to the segment splitter, which consists of dividing the original signal into L segments:

$$\mathbf{p}(t) = [p_1(t), \ldots, p_M(t)]^T, \ t = 1, \ldots, L \tag{17}$$

where M is the segment length. The plain signal contains $L \times M$ samples, it is split in L segments of M samples, the M samples form the $M \times 1$ vector $\mathbf{p}(t)$. These segments are used in the encryption process (the subspace-based encryption block) to obtain the following encrypted signal $\mathbf{x}(t)$:

$$\mathbf{x}(t) = \mathbf{A}(t)\mathbf{p}(t) + \beta \mathbf{P}^{\perp}_{A(t)} \mathbf{B}(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{p}(t))] \tag{18}$$

where $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are $(M + 1) \times M$ and $(M + 1) \times M$ full rank key matrices, respectively. The introduction of matrices $\mathbf{A}(t)$ and $\mathbf{B}(t)$ is motivated by the task of increasing the key space that would be needed for cryptanalysis. Note that the key matrices are generated for each vector $\mathbf{x}(t)$. This property makes any estimation of the signal subspace impossible from only one snapshot. These matrices can be generated by a Pseudorandom Number Generator (PRNG) with a secret seed that serves as the secret key. β is a factor that controls the signal-to-noise ratio. This factor (β) should be chosen as large as possible in order to provide very low Signal-to-Noise Ration (SNR), $\mathbf{g}(.)$ is a component-wise nonlinear function that verifies

$$\mathbf{g}(0) = 0. \tag{19}$$

$\mathbf{k}(t)$ is a random $M \times 1$ key signal vector generated by any robust key signal generator and $\odot$ denotes the Hadamard operator. $\mathbf{P}^{\perp}_{A(t)}$ is the projector on the orthogonal subspace to the one spanned by the columns of the key matrix $\mathbf{A}(t)$. The latter is refereed herein to as the key subspace. The projector $\mathbf{P}^{\perp}_{A(t)}$ is given by

$$\mathbf{P}^{\perp}_{A(t)} = \mathbf{I} - \mathbf{P}_{A(t)} = \mathbf{I} - \mathbf{A}(t)(\mathbf{A}(t)^H \mathbf{A}(t))^{-1} \mathbf{A}(t)^H \tag{20}$$

where $\mathbf{P}_{A(t)}$ is the orthogonal projector on the key subspace, and $(.)^H$ and $\mathbf{I}$ denote the Hermitian operator and the identity matrix, respectively. For the purpose of robustness evaluation, we use in the sequel the following component-wise nonlinear function:

$$g(v) = \frac{v}{\sqrt{1 + v^2}} \tag{21}$$

that verifies condition (19).

4.2 Decryption

On the receiver hand, the encrypted data vector is first projected on the corresponding key subspace; this is done by the following operation:

$$\mathbf{x}_p(t) = \mathbf{P}_{A(t)}\mathbf{x}(t) \tag{22}$$

where $\mathbf{x}_p(t)$ is the obtained projected data. Since the projectors $\mathbf{P}_{A(t)}$ and $\mathbf{P}^{\perp}_{A(t)}$ are orthogonal (i.e., $\mathbf{P}_{A(t)}\mathbf{P}^{\perp}_{A(t)} = \mathbf{0}$), the above projection leads to the following result:

$$\mathbf{x}_p(t) = \mathbf{A}(t)\mathbf{p}(t) \tag{23}$$

and the original plain-text (the decrypted signal) is obtained by using the key matrix $\mathbf{A}(t)$:

$$\mathbf{p}(t) = (\mathbf{A}(t))^{\sharp}\mathbf{x}_p(t) \tag{24}$$

where $(.)^{\sharp}$ denotes the pseudoinverse operator. Note that in the above recovery procedure, one does not need to know the key signals $\mathbf{k}(t)$ neither the matrix $\mathbf{B}(t)$. The $\mathbf{k}(t)$ in Eq. (18) is not the same as that used in the blind source separation-based encryption scheme. The difference is in the way of its use. In blind source separation-based encryption scheme, $\mathbf{k}(t)$ is included in the transmitted source vector in order to generate an underdetermined Blind Source Separation source (BSS) problem. In the subspace-based method, $\mathbf{k}(t)$ is used in conjuncture with a nonlinearity of the data as an additive perturbation term that also generates an underdetermined Blind Source Separation source (BSS) problem. The dimension of $\mathbf{k}(t)$ is $M \times 1$. $\mathbf{k}(t)$ is generated by any pseudorandom or random generator. There is no specific range for values of $k(t)$. It depends on the chosen key generator. Of course, the randomness quality of the key signals of any encryption system can affect the encryption results. In our case, this is also true. However, the randomness quality is guaranteed by the pseudorandom or random generator used in the encryption system (see [12, 33–35]). For simplicity, the MATLAB generic instruction "random" has been used. Several pseudorandom and random generators exist, both in software and hardware forms, and any generator which fulfills the randomness criteria largely is described in the literature (e.g., NIST tests of randomness, Maurer test) [36, 37] can be used to generate $\mathbf{k}(t)$. However, the issue of studying and evaluating the randomness of the key generators, both pseudorandom and random, is over the scope of this chapter.

The factor β in Eq. (18) should be chosen as large as possible in order to provide very low SNR. For such chosen values that should lead to secure encrypted data, we can not get small eigenvalues for this subspace. If one set $\beta = 0$, this means that we have no encryption according to the subspace-based scheme. Even if one can estimate $\mathbf{A}$ from the encrypted data $\mathbf{x}$, the estimate $\hat{\mathbf{A}}$ will be given with some estimation error say $\Delta\mathbf{A}$:

$$\hat{\mathbf{A}} = \mathbf{A} + \Delta\mathbf{A}.$$

Note that the existing correlation between the two terms of Eq. (18) will increase the estimation error if one find a way to estimate **A** or its subspace. Besides this property, several tests have been conducted to measure the sensitivity of the subspace-based scheme even to very small matrix mismatch. Results of these tests are presented in Sect. 7.

5 Iterative Subspace-Based Encryption

The iteration of an encryption scheme for a number of rounds is usually applied on cryptosystems to enhance their security characteristics and hence, strengthen their resistance to cryptanalysis attacks. At this level, the subspace-based encryption scheme described above constitutes one round. The output of the first round is re-injected as the input for the second round and so on. The output of the last round represents the output of the whole encryption scheme, i.e., the iterative subspace-based encryption scheme. Figure 3 shows the block diagram of the proposed iterative subspace-based encryption scheme.

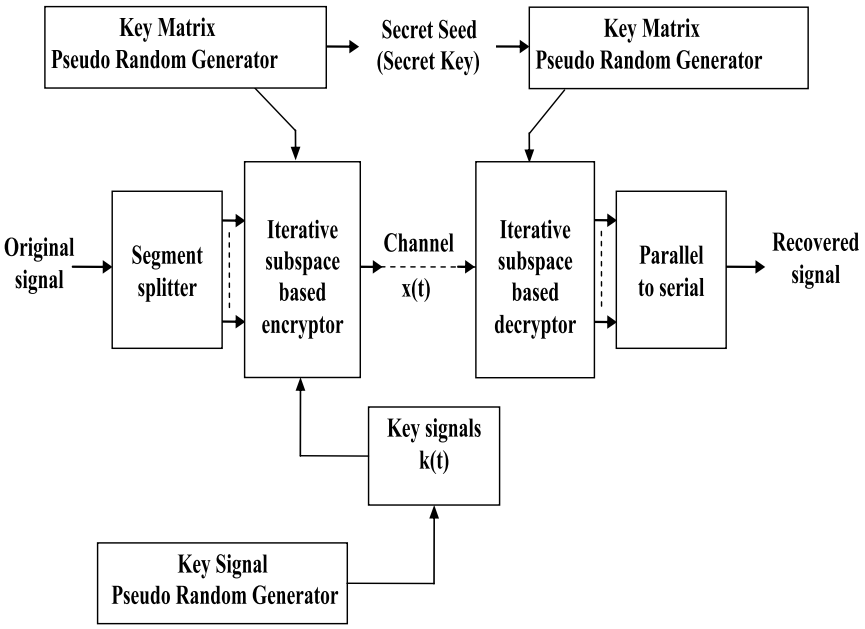


Fig. 3 Block diagram of the iterative subspace-based encryption

5.1 Encryption

The splitted segments of Eq. (17) are used in the iterative encryption process to obtain the following encrypted signal:

$$\mathbf{x}_n(t) = \mathbf{A}_n(t)\mathbf{x}_{(n-1)}(t) + \beta \mathbf{P}^{\perp}_{A_n(t)}\mathbf{B}_n(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{x}_{(n-1)}(t))], \tag{25}$$

where $\mathbf{x}_n(t)$ and $\mathbf{x}_{(n-1)}(t)$ denote the nth and $(n-1)$th encrypted segments with $n \geq 1$, respectively, and $\mathbf{x}(0) = \mathbf{p}(t)$ is the plain-text. $\mathbf{A}_n(t)$ and $\mathbf{B}_n(t)$ are $(M+1) \times M$ full rank key matrices, respectively. Note that the encryption process described in Eq. (25) is performed on several iterations.

5.2 Decryption

Once the cipher-text is obtained, the decryption procedure could be achieved by projecting the last encrypted segment $\mathbf{x}_{n(t)}$ as described by the following equation:

$$\mathbf{x}_{p,n}(t) = \mathbf{P}_{A_n(t)}\mathbf{x}_{n(t)} \tag{26}$$

where $\mathbf{x}_{p,n}(t)$ is the obtained projected data. Since the projectors $\mathbf{P}_{A_n(t)}$ and $\mathbf{P}^{\perp}_{A_n(t)}$ are orthogonal (i.e., $\mathbf{P}_{A_n(t)}\mathbf{P}^{\perp}_{A_n(t)} = \mathbf{0}$), the above projection leads to the following result:

$$\mathbf{x}_{p,n}(t) = \mathbf{A}_n(t)\mathbf{x}_{(n-1)}(t). \tag{27}$$

The decrypted signal at iteration $n-1$, is then obtained by

$$\mathbf{x}_{(n-1)}(t) = (\mathbf{A}_n(t))^{\sharp}\mathbf{x}_{p,n}(t), \tag{28}$$

where $(.)^{\sharp}$ denotes the pseudoinverse operator. The above equations are performed iteratively till restituting the original plain-text.

6 Cryptographic Robustness of the Subspace-Based Encryption Systems

The subspace-based encryption schemes are evaluated to assess their robustness, from a cryptographic point of view, and their quality of recovering the original signal (plain-text). The cryptographic robustness assessment approach is cryptanalysis oriented in the sense that cryptanalysis attacks are applied on the proposed subspace-based encryption schemes. Results of these cryptanalysis attacks are used to make a comparison with the BSS-based encryption scheme.

6.1 Interpretation of the Subspace-Based Encryption in Terms of Confusion and Diffusion Requirements

In the design of most published cryptographic systems, two important principles are present in the designer's mind: Confusion and Diffusion. Confusion is based on the idea of obscuring the relationship between plain-text, cipher-text, and keys. This is done by mixing linearity and nonlinearity [38]. Diffusion is the other important principle of cryptographic system design and is based on the idea that every bit of the cipher-text should depend on every bit of the plain-text and every bit of the key. This ensures that the statistics of the plain-text are dissipated within the cipher-text so that an attacker cannot predict the plain-text that corresponds to a particular cipher-text, even after observing a number of "similar" plain-texts and their corresponding cipher-texts [38]. Generally, in most of the published cryptographic systems, substitution and permutation are the main two operations applied, both or separately, on plain-texts in order to ensure confusion and diffusion. While the terminology (substitution and permutation) used nowadays is roughly the same since centuries and the objective is to make the cipher-text the most complex and unintelligible, the confusion and diffusion terms are relatively recent. Of course, the approaches and techniques have seen a huge development during the long history of cryptology to ensure confusion and diffusion. As an example, the security of Advanced Encryption Standard (AES) [39], the most known and recent cryptographic standard, is mainly based on the robustness of S-boxes, the "Substitution" boxes. Diffusion in the AES *SP*-network is achieved by a linear transformation [40]. On another hand, some cryptographic systems use other approaches and techniques to guarantee a high degree of confusion and diffusion. The subspace-based encryption system are in this category of cryptographic systems. There is no permutation or substitution in the known sense of the terms, rather there is a new approach based on subspace concept to guarantee the same security objectives targeted by substitution and permutation, i.e., confusion and diffusion. In the subspace-based encryption system, confusion is achieved by the linearity and nonlinearity that obscure the relationship between the plain-text, the cipher-text, and the key. Furthermore, we see from Eq. (29) that each value of the cipher-text $\mathbf{x}(t)$ depends on each value of the plain-text and the key what ensures the diffusion requirement. Next, the robustness, from a security point of view, of the proposed subspace-based cryptographic system is analyzed and evaluated.

6.2 Cipher-Text-Only Attack

This is the most known and realistic attack since it does not require more than the availability of cipher-texts, which can be obtained by a system similar to the system which is under attack or even by interception of cipher-texts. Generally, communications are using public infrastructures and protocols (e.g., telephone networks, internet) and hence, could be intercepted.

6.2.1 Sensitivity to $\mathbf{P}_{A(t)}$

To evaluate the sensitivity of a cryptosystem to its encryption key, two different keys are used for the same plain-text. This evaluation is related to the obtained difference between the two encrypted signals. For a robust cryptosystem, this difference should be completely different for any pair of keys, even with minimal value difference under the considered finite precision [9, 31]. From a cryptographic point of view, a robust cryptosystem should have a high sensitivity to the encryption key [9, 31]. In the case of a very low sensitivity, a mismatched key would approximately recover the plain-text. Next, it is shown that the subspace-based encryption scheme is very sensitive to projection mismatch.

Subspace-based encryption scheme
Let us first rewrite for ease of use the encryption equation (18) as

$$\mathbf{x}(t) = \mathbf{y}_p(t) + \beta \mathbf{z}(t), \tag{29}$$

where $\mathbf{y}_p(t) = \mathbf{A}(t)\mathbf{p}(t)$ and $\mathbf{z}(t) = \mathbf{P}^{\perp}_{A(t)}\mathbf{B}(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{p}(t))]$. Consider the following mismatched projector

$$\hat{\mathbf{P}}_{A(t)} = \mathbf{P}_{A(t)} + \varepsilon \mathbf{I} \tag{30}$$

with ε a finite precision value and $\mathbf{I}$ a $(M+1) \times (M+1)$ identity matrix. Using the mismatched projector $\hat{\mathbf{P}}_{A(t)}$ for decryption, one gets the following:

$$\begin{aligned} \hat{\mathbf{P}}_{A(t)}\mathbf{x}(t) &= (\mathbf{P}_{A(t)} + \varepsilon \mathbf{I})\mathbf{y}_p(t) + \beta(\mathbf{P}_{A(t)} + \varepsilon \mathbf{I})\mathbf{z}(t) \\ &= (1+\varepsilon)\mathbf{y}_p(t) + \beta\varepsilon \mathbf{z}(t). \end{aligned} \tag{31}$$

In Eq. (31), we have used the fact that

$$\mathbf{P}_{A(t)}\mathbf{y}_p(t) = \mathbf{y}_p(t) \text{ and } \mathbf{P}_{A(t)}\mathbf{z}(t) = 0.$$

Note from the same Eq. (31) that the decrypted data by the mismatched projector is still encrypted according to the proposed encryption equation (18). By choosing $\beta = O(\frac{1}{\varepsilon})$, Eq. (31) shows that, even for a very small value of ε, there is a very significant difference between the decryption results obtained by the actual projector $\mathbf{P}_{A(t)}$ and its mismatched version $\hat{\mathbf{P}}_{A(t)}$. This means that the proposed subspace-based encryption scheme is very sensitive to projector mismatch. Hence, it verifies an important principle of cryptographic robustness that is the high sensitivity to secret parameter mismatch. This high sensitivity is checked by using the following experimental procedure [9]:

- *Step 1*: For a randomly generated projector and keys $(\mathbf{P}_{A(t)}, \mathbf{k}(t))$, compute the cipher-text $\mathbf{x}(t)$ corresponding to a plain-text $\mathbf{p}(t)$.

- *Step 2*: With a mismatched projector $\mathbf{P}_{A(t)} + \varepsilon\mathbf{I}$, decrypt $\mathbf{x}(t)$ to get $\hat{\mathbf{p}}(t)$, an estimated version of $\mathbf{p}(t)$, where $\varepsilon \in [0, 1]$.

Detailed results and discussion of this experiment are shown in Sect. 7.

Iterative subspace-based encryption scheme
For the iterative subspace-based encryption scheme, let us rewrite the encryption equation:

$$\mathbf{x}_n(t) = \mathbf{A}_n(t)\mathbf{x}_{(n-1)}(t) + \beta\mathbf{P}^{\perp}_{A_n(t)}\mathbf{B}_n(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{x}_{(n-1)}(t))] \tag{32}$$

where $\mathbf{x}_n(t)$ and $\mathbf{x}_{(n-1)}(t)$ denote the nth and $(n-1)$th encrypted segments. $n \geq 1$ and $\mathbf{x}(0) = \mathbf{p}(t)$, the plain-text as

$$\mathbf{x}_n(t) = \mathbf{y}_{(p,n)}(t) + \beta\mathbf{z}_n(t), \tag{33}$$

where $\mathbf{y}_{(p,n)}(t) = \mathbf{A}_n(t)\mathbf{x}_{(n-1)}(t)$ and $\mathbf{z}_n(t) = \mathbf{P}^{\perp}_{A_n(t)}\mathbf{B}_n(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{x}_{(n-1)}(t))]$. If we consider the following mismatched projector

$$\hat{\mathbf{P}}_{\mathbf{A}_n(t)} = \mathbf{P}_{\mathbf{A}_n(t)} + \varepsilon\mathbf{I} \tag{34}$$

with ε a finite precision value and $\mathbf{I}$ a $(M+1) \times (M+1)$ identity matrix. The decrypted data $\mathbf{x}_{(p,n)}$ using the mismatched projector is then given by $\hat{\mathbf{P}}_{\mathbf{A}_n(t)}\mathbf{x}_n(t)$. One gets

$$\begin{aligned}\hat{\mathbf{P}}_{\mathbf{A}_n(t)}\mathbf{x}_n(t) &= (\mathbf{P}_{\mathbf{A}_n(t)} + \varepsilon\mathbf{I})\mathbf{y}_{(p,n)}(t) + \beta(\mathbf{P}_{\mathbf{A}_n(t)} + \varepsilon\mathbf{I})\mathbf{z}_n(t) \\ &= (1+\varepsilon)\mathbf{y}_{(p,n)}(t) + \beta\varepsilon\mathbf{z}_n(t)\end{aligned} \tag{35}$$

Note from Eq. (35) that the decrypted data by the mismatched projector is still encrypted according to the proposed encryption equation (18). By choosing $\beta = O(\frac{1}{\varepsilon})$, Eq. (35) shows that, even for a very small value of ε, there is a very significant difference between the decryption results obtained by the actual projector $\mathbf{P}_{\mathbf{A}_n(t)}$ and its mismatched version $\hat{\mathbf{P}}_{\mathbf{A}_n(t)}$. Note that the iteration on n makes an accumulation on the initial mismatch on the projector and hence makes the encryption more sensitive to projector mismatch. Adopting the same methodology applied for the subspace-based encryption scheme to check the high sensitivity of iterative subspace-based encryption scheme to key mismatch, the following procedure is applied:

- Step 1: Generate random projector and keys $(\mathbf{P}_{\mathbf{A}_n(t)}, \mathbf{k}(t))$, then compute the ciphertext $\mathbf{x}_n(t)$ corresponding to a plain-text $\mathbf{p}(t)$.
- Step 2: Using a mismatched projector $\mathbf{P}_{\mathbf{A}_n(t)} + \varepsilon\mathbf{I}$, decrypt $\mathbf{x}_n(t)$ to get $\hat{\mathbf{p}}(t)$, an estimated version of $\mathbf{p}(t)$, where $\varepsilon \in [0, 1]$.

Detailed results and discussion of this experiment are shown in Sect. 7.

6.2.2 Sensitivity to $\mathbf{k}(t)$

The subspace-based encryption scheme is also very sensitive to the key signals. This is due to the same reason of its high sensitivity to projector $\mathbf{P}_{A(t)}$. Furthermore, the key signals $\mathbf{k}(t)$ are generated randomly for each plain-text through the random or the pseudorandom generator as shown in Fig. 2. Let us rewrite the encryption equation (18) for the subspace-based encryption scheme as follows:

$$\mathbf{x}(t) = \mathbf{A}(t)\mathbf{p}(t) + \beta \mathbf{P}^{\perp}_{A(t)} \mathbf{B}(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{p}(t))], \tag{36}$$

where $\mathbf{A}(t)$ and $\mathbf{B}(t)$ are $(M+1) \times M$ and $(M+1) \times M$ full rank key matrices, respectively. $\mathbf{k}(t)$ is generated for each vector $\mathbf{x}(t)$. Consider now a second operation of encrypting the same plain-text $\mathbf{p}(t)$ using the same key matrix $\mathbf{A}(t)$. An expected result of this encryption operation should be the same cipher-text $\mathbf{x}(t)$ obtained as an output of the Eq. (36). However, this is not the case in our proposed subspace-based encryption scheme. The obtained cipher-text is given by the following equation:

$$\mathbf{x}_1(t) = \mathbf{A}(t)\mathbf{p}(t) + \beta \mathbf{P}^{\perp}_{A(t)} \mathbf{B}(t)[\mathbf{k}_1(t) \odot \mathbf{g}(\mathbf{p}(t))], \tag{37}$$

where the cipher-text $\mathbf{x}_1(t)$ is different from $\mathbf{x}(t)$ because the key signal $\mathbf{k}_1(t)$ is totally different from the key signal $\mathbf{k}(t)$ used in the first encryption. This is due to the random (or pseudorandom) generator used to generate the key signals that generates each time a different sequence of keys. More generally, the cipher-texts obtained from the use of the same plain-text and the same key matrix in a subspace-based encryption scheme are always different. This is an important feature which has an impact on the resistance of the proposed subspace encryption scheme to cipher-text only attack. Actually, a cryptanalyst gathering a set of let us say N cipher-texts has no information about the number of the corresponding plain-texts. This number could be the same number of cipher-texts or less. This uncertainty about the number of the corresponding plain-texts provides an additive level of resistance to this class of cryptanalysis attack.

6.2.3 Sensitivity to Plain-Text

A robust cryptosystem is also required to be very sensitive to plain-text. Hence, the cipher-texts of two slightly different plain-texts should be significantly different [9, 31]. The subspace-based encryption scheme matches well this property. The subspace-based encryption of two slightly different plain-texts $p^{(1)}(t)$ and $p^{(2)}(t)$ by the same key leads to a significant sensitivity to the slight difference. It is observed for the speech application, that a person can not distinguish the difference between the two slightly different plain-texts, however, their cipher-texts are totally different.

6.3 *Differential Attack*

In a differential attack, the encryption of at least two plain-texts by two identical key signals is assumed [9]. However, the key space of the used pseudorandom generator, should be large enough to avoid the occurrence of two identical keys. Moreover, even when assuming that two identical key signals have been used to encrypt two plain-texts through the subspace-based encryption scheme, the differential attack cannot be realized. This is mainly due to the difficulty of solving nonlinear equations.

Subspace-based encryption scheme
Let us assume that two plain-texts $\mathbf{p}^{(1)}(t)$ and $\mathbf{p}^{(2)}(t)$ are encrypted using the same key parameters $(\mathbf{P}_{A(t)}, \mathbf{k}(t))$. From Eq. (18), one has

$$\mathbf{x}^{(1)}(t) = \mathbf{A}(t)\mathbf{p}^{(1)}(t) + \beta \mathbf{P}^{\perp}_{A(t)}\mathbf{B}(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{p}^{(1)}(t))] \tag{38}$$

and

$$\mathbf{x}^{(2)}(t) = \mathbf{A}(t)\mathbf{p}^{(2)}(t) + \beta \mathbf{P}^{\perp}_{A(t)}\mathbf{B}(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{p}^{(2)}(t))]. \tag{39}$$

Combining Eqs. (38) and (39), leads to

$$\Delta_{\mathbf{x}}(t) = \mathbf{A}(t)\Delta_{\mathbf{p}}(t) + \beta \mathbf{P}^{\perp}_{A(t)}\mathbf{B}(t)[\mathbf{k}(t) \odot [\mathbf{g}(\mathbf{p}^{(1)}(t)) - \mathbf{g}(\mathbf{p}^{(2)}(t))]], \tag{40}$$

where $\Delta_{\mathbf{x}}(t)$ is the cipher-text differential described as

$$\Delta_{\mathbf{x}}(t) = \mathbf{x}^{(1)}(t) - \mathbf{x}^{(2)}(t) \tag{41}$$

and $\Delta_{\mathbf{p}}(t)$ is the plain-text differential described as

$$\Delta_{\mathbf{p}}(t) = \mathbf{p}^{(1)}(t) - \mathbf{p}^{(2)}(t). \tag{42}$$

Note that even if the same key is used to obtain the cipher-texts $\mathbf{x}^{(1)}(t)$ and $\mathbf{x}^{(2)}(t)$, the additive subspace perturbation term is still present in Eq. (40). Moreover, the plain-text differential $\Delta_{\mathbf{p}}(t)$ cannot be computed, because of the permanent presence of the terms $\mathbf{p}^{(1)}(t)$ and $\mathbf{p}^{(2)}(t)$ in Eq. (40). This is due, as mentioned in Sect. 4, to the existing correlation between the additive subspace perturbation term and the plain-text term of the proposed encryption equation (18).

Iterative subspace-based encryption scheme
By assuming that two plain-texts $\mathbf{p}^{(1)}(t)$ and $\mathbf{p}^{(2)}(t)$ are encrypted using the same key parameters $(\mathbf{P}_{\mathbf{A}_n(t)}, \mathbf{k}(t))$. From Eq. (25), one has

$$\mathbf{x}^{(1)}_n(t) = \mathbf{A}_n(t)\mathbf{x}^{(1)}_{(n-1)}(t) + \beta \mathbf{P}^{\perp}_{\mathbf{A}_n(t)}\mathbf{B}_n(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{x}^{(1)}_{(n-1)}(t))], \tag{43}$$

where $\mathbf{x}^{(1)}_n(t)$ and $\mathbf{x}^{(1)}_{(n-1)}$ denote the nth and $(n-1)$th encrypted segments of the first plain-text with $n \geq 1$, respectively, $\mathbf{x}^{(1)}(0) = \mathbf{p}^{(1)}(t)$ is the first plain-text and

$$\mathbf{x}_n^{(2)}(t) = \mathbf{A}_n(t)\mathbf{x}_{(n-1)}^{(2)}(t) + \beta \mathbf{P}_{\mathbf{A}_n(t)}^{\perp}\mathbf{B}_n(t)[\mathbf{k}(t) \odot \mathbf{g}(\mathbf{x}_{(n-1)}^{(2)}(t))], \tag{44}$$

where $\mathbf{x}_n^{(2)}(t)$ and $\mathbf{x}_{(n-1)}^{(2)}$ denote the nth and $(n-1)$th encrypted segments of the second plain-text with $n \geq 1$, respectively, and $\mathbf{x}^{(2)}(0) = \mathbf{p}^{(2)}(t)$ is the second plain-text. Combining Eqs. (43) and (44), one gets

$$\begin{aligned} \Delta_{\mathbf{x}_n}(t) &= \mathbf{A}_n(t)\Delta_{\mathbf{x}_{(n-1)}}(t) \\ &+ \beta \mathbf{P}_{\mathbf{A}_n(t)}^{\perp}\mathbf{B}_n(t)[\mathbf{k}(t) \odot [\mathbf{g}(\mathbf{x}_{(n-1)}^{(1)}(t)) - \mathbf{g}(\mathbf{x}_{(n-1)}^{(2)}(t))]], \end{aligned} \tag{45}$$

where $\Delta_{\mathbf{x}_n(t)}$ and $\Delta_{\mathbf{x}_{(n-1)}}$ are the cipher-text differentials described as

$$\Delta_{\mathbf{x}_n(t)} = \mathbf{x}_n^{(1)}(t) - \mathbf{x}_n^{(2)}(t) \tag{46}$$

and

$$\Delta_{\mathbf{x}_{(n-1)}}(t) = \mathbf{x}_{(n-1)}^{(1)}(t) - \mathbf{x}_{(n-1)}^{(2)}(t), \tag{47}$$

where

$$\begin{aligned} \Delta_{\mathbf{x}_0(t)} &= \mathbf{x}_0^{(1)}(t) - \mathbf{x}_0^{(2)}(t) \\ &= \mathbf{p}^{(1)}(t) - \mathbf{p}^{(2)}(t) \\ &= \Delta_{\mathbf{p}}(t) \end{aligned} \tag{48}$$

One can see that even if the same key is used to obtain the cipher-texts $\mathbf{x}_n^{(1)}(t)$ and $\mathbf{x}_n^{(2)}(t)$, the additive subspace perturbation term is still present in Eq. (45). Moreover, the plain-text differential $\Delta_{\mathbf{p}}(t)$ cannot be computed because of the permanent presence of the terms $\mathbf{p}^{(1)}(t)$ and $\mathbf{p}^{(2)}(t)$ in Eq. (45). This is due again to the existing correlation between the additive subspace perturbation term and the plain-text term of the proposed encryption equation (25).

7 Application and Performance Evaluation

This section presents the results of experiments conducted on the subspace-based encryption method and its iterative version. The latter assess the proposed schemes by evaluating aspects related to both security robustness from a cryptographic point of view and quality of reconstruction of plain-texts at the decryption level. For security assessment purpose, the evaluation approach adopts some cryptanalysis attacks. For quality assessment, the evaluation process uses both subjective and objective measurements. The tests and experimentations were conducted on speech signals and images.

7.1 Application to Speech Encryption

7.1.1 Security Robustness Evaluation

In practice, ε the finite precision value that would be used in the cryptanalysis by exhaustive search, varies usually from 0.1 to 0.01. If ε is chosen too small, the key space becomes huge and the cryptanalysis by exhaustive search becomes impracticable. The decryption requires a matrix pseudoinversion. This might be expensive (especially if M is large) and might result in numerical problems if A(t) is ill-conditioned. However, the numerical problems that could arise from the possible ill-conditioning of matrix A(t) depend on the quality of the random or pseudorandom generator. In our experiments, M is chosen equal to 4, we have also run experiments with $M = 2$ and there were no significant effects on the encryption performance.

7.1.2 Quality Performance Analysis

For the objective evaluation, the Signal-to-Noise Ratio (SNR) in dB of each original segment in both the encrypted and decrypted segments is considered. The subspace-based encryption is applied to a speech record of a child singing a song entitled "let's laugh together". The obtained results are shown in Table 1. The signal is sampled at 22.05 kHz and encoded by 16 bits/sample. From Table 1, it appears clearly that the original segments are well hidden by the subspace-based approach while the decrypted segments are recovered with a very high SNR ensuring an excellent voice quality in the case of speech encryption. Note that the encrypted segments have a very low SNR.

We have also conducted a subjective evaluation through a listening test that uses the subjective Degradation Category Rating (DCR). A 5-point scale has been used: Degradation is very annoying (1), Degradation is annoying (2), Degradation is slightly annoying (3), Degradation is audible but not annoying (4), and Degradation is inaudible (5) [5, 41]. The Degradation Mean Opinion Score (DMOS) is obtained as the mean value of the listener's appreciation. Twenty listeners, ten male and ten female, were selected to give their scoring after hearing the original and the

Table 1 SNR(dB) of four original speech segments in four encrypted segments and four decrypted segments

	$x_1(t)$	$x_2(t)$	$x_3(t)$	$x_4(t)$	$x_p(t)$
$p_1(t)$	−204.28	−189.76	−183.23	−206.62	404.25
$p_2(t)$	−208.66	−193.14	−186.61	−210.00	361.62
$p_3(t)$	−216.20	−200.68	−194.14	−217.54	341.09
$p_4(t)$	−207.39	−191.87	−185.34	−208.73	352.03

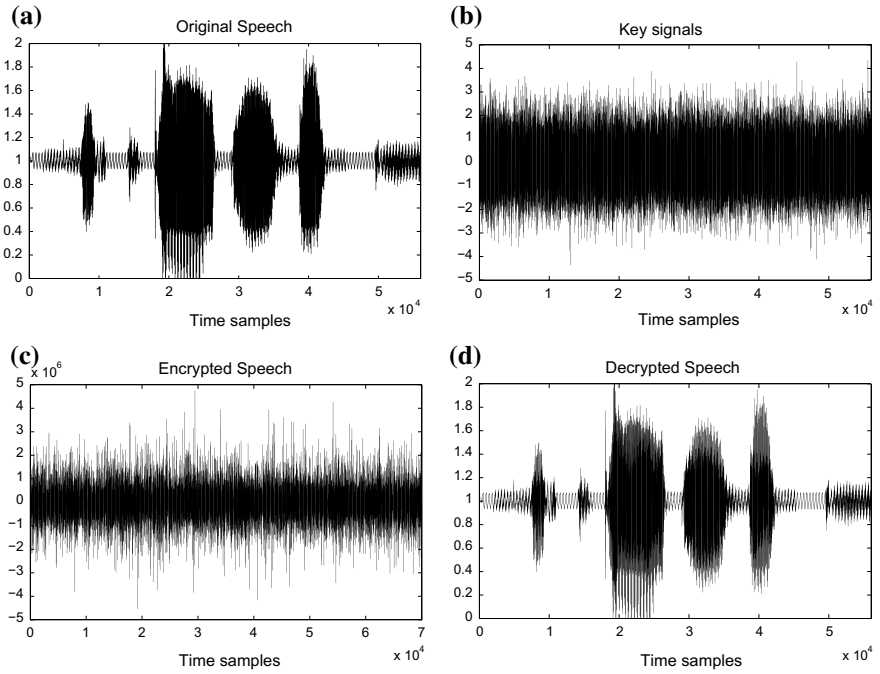


Fig. 4 An example of subspace-based speech encryption, **a** Original speech, **b** Key signals, **c** Encrypted speech, **d** Decrypted speech

decrypted speeches. The obtained DMOS was 5. Thus, the listeners have approved the excellent voice quality.

Figure 4 shows the results obtained from applying subspace- based encryption on the speech record described above with a factor β equal to 10^6. Figure 4a shows the original signal, whereas Fig. 4b shows the key signals $\mathbf{k}(t)$ used during encryption. Figure 4c shows the signal encrypted according to Eq. (18) with a segment length $M = 4$. Note also that the encrypted signal has more samples than the original one, actually L samples more where L is the segment number. This sample excess comes from the fact that the dimension of the key matrices $\mathbf{A}(t)$ is $(M + 1) \times M$. After decryption with the proposed subspace method, the recovered signal is shown in Fig. 4d. As one can see from this figure, there is no visual difference between the original speech and the decrypted one. Figure 5 shows results obtained from applying iterative subspace encryption on the same speech file used in subspace-based encryption with a factor β equal to 10^6. Figure 5a shows the original signal, whereas Fig. 5b and c show the key signals $\mathbf{k}(t)$ used during encryption and the obtained encrypted signal, respectively, according to Eq. (25) with a segment length $M = 4$ and 2-rounds encryption. Note that, as mentioned in the subspace-based encryption, the encrypted signal has also more samples than the original one. Figure 5d shows the recovered signal after decryption. One can see the similarity between the original

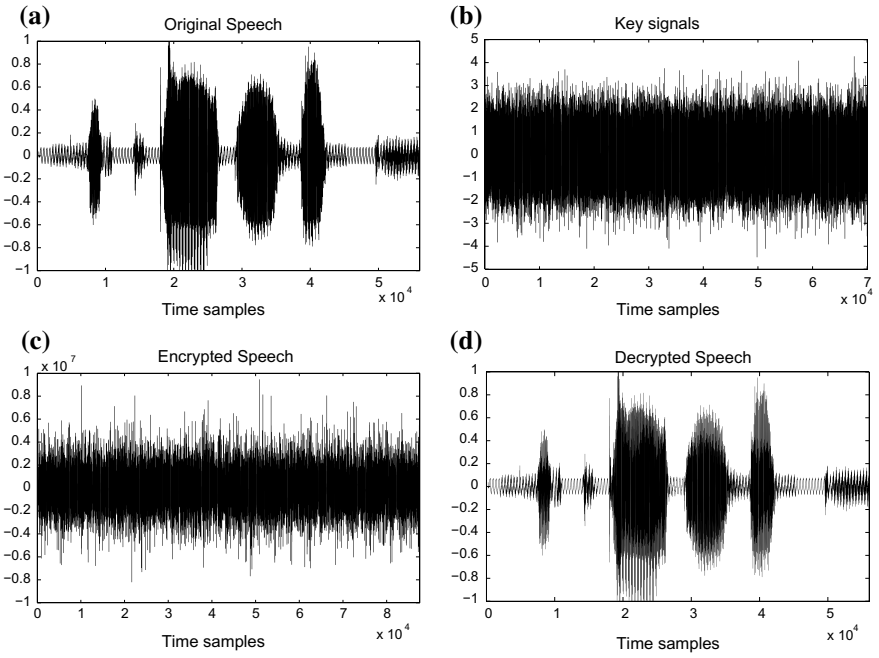


Fig. 5 An example of iterative subspace-based speech encryption with 2 iterations, **a** Original speech, **b** Key signals, **c** Encrypted speech, **d** Decrypted speech

speech and the decrypted one. In order to assess the impact of the iterative rounds of the subspace-based encryption, a comparison is made between 1-round and several ones. A comparison is conducted between the cipher-texts and their corresponding decrypted signals obtained from the use of subspace-based encryption and its iterative version for different iterations, Fig. 6 shows the results for 1-round, 2-rounds, 3-rounds, and 4-rounds encryption, respectively. One can see that the amplitude level of the encrypted signal increases proportionally to the number of iterations. Actually, the amplitude level is multiplied by a factor 2 when the number of iterations increases by 1. At the decryption side, the original signal is recovered and no visual difference is found between the recovered signals of the different iterations.

Figures 7 and 8 show a comparison in terms of sensitivity levels to plain-text mismatches of 0.1 and 0.01, respectively, between iterative subspace-based encryption schemes for different iterations: 1-round, 2-rounds, 3-rounds, and 4-rounds encryption. One can see that, for the same level of plain-text mismatch, the sensitivity of the subspace-based encryption scheme, revealed by the cipher-text difference level, increases when the number of iterations rises. On the other side, even when the plain-text mismatch level decreases (from 0.1 to 0.01), the sensitivity decreases but remains at high levels, with a cipher-text difference varying roughly between 10^4 and 10^7.

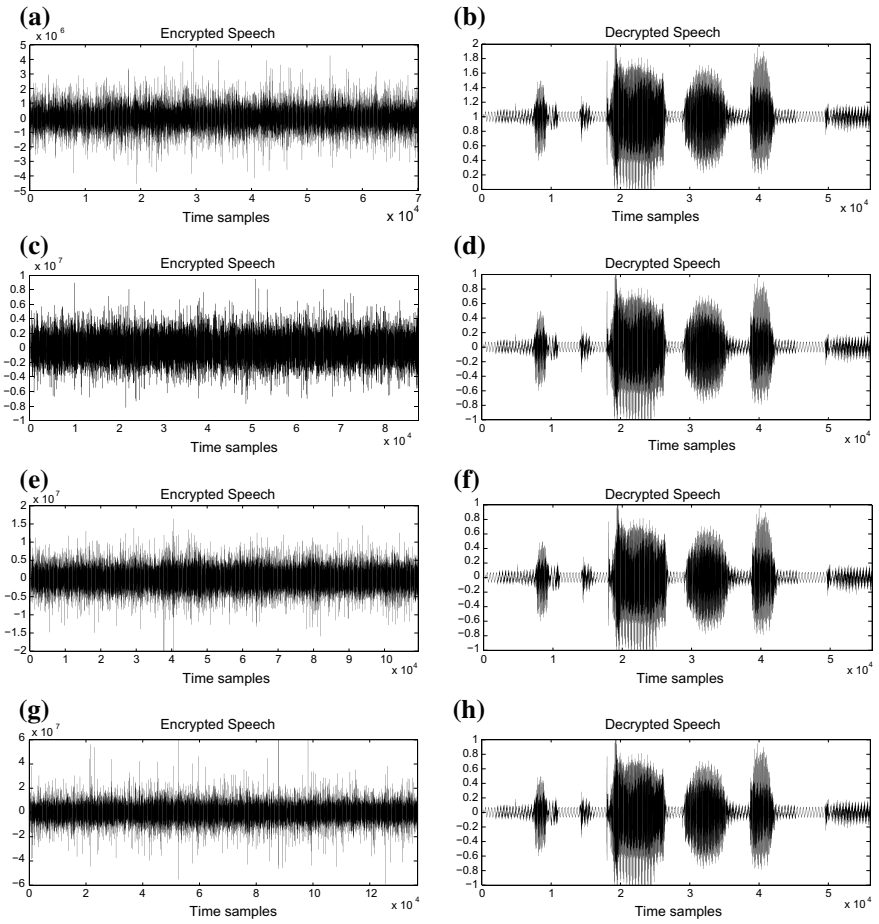


Fig. 6 A comparison between cipher-texts and recovered signals for the subspace-based speech encryption and its iterative version, **a**–**b** 1-round, **c**–**d** 2-rounds, **e**–**f** 3-rounds, **g**–**h** 4-rounds

In Figure 9a, the experimental relationship between the recovery error and the value of mismatch level ε is plotted for different iterations when the iterative subspace encryption is used. The value of the factor β is equal to 10^6 and the values of ε vary from 10^{-3} to 1. For the smallest value of ε used during experimentations, one can see that the corresponding recovery error expressed in terms of Mean Absolute Error (MAE) is high (10^4) for a 2-rounds encryption. However, for the same smallest value of ε, the recovery error rises to 10^5 for a 6-rounds encryption.

The second experimental relationship, between the recovery error and the value of the factor β for the iterative encryption scheme is shown in Fig. 9b. One can see that there is a linear relationship between the factor β and the recovery error. For the

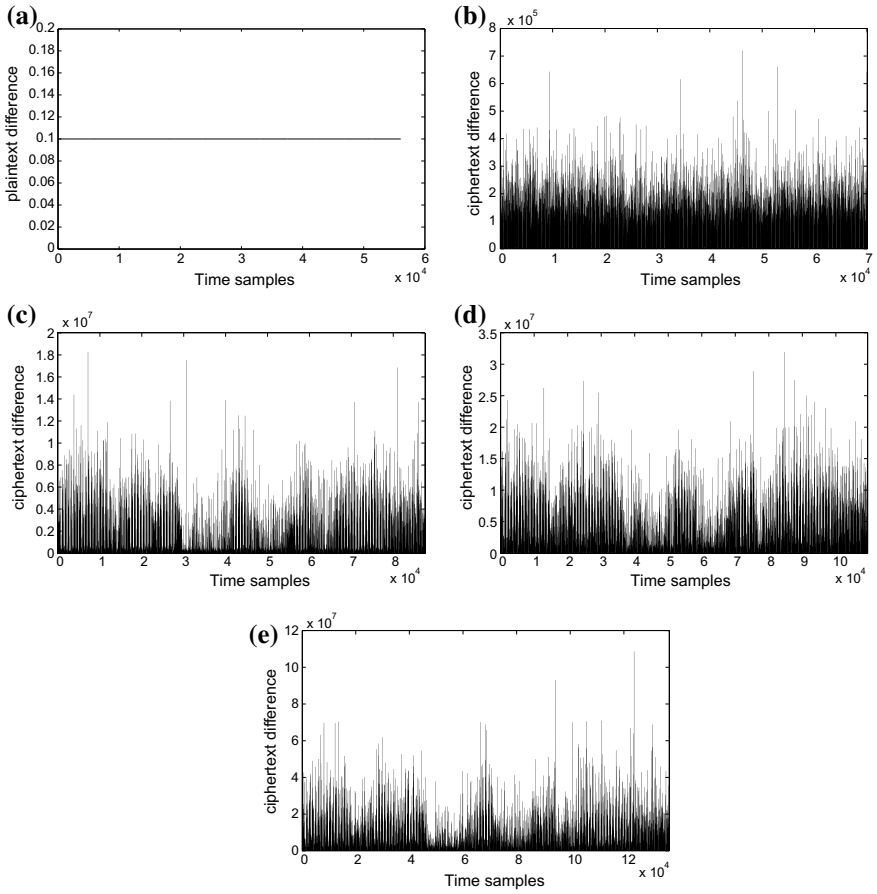


Fig. 7 A comparison in terms of sensitivity levels to a 0.1 plain-text mismatch between iterative subspace-based encryption schemes for different iterations, **a** plain-text mismatch, cipher-text difference for: **b** 1-round, **c** 2-rounds, **d** 3-rounds, **e** 4-rounds

same value of β, let us say 10^6, the recovery error varies from 10^4 to 10^5 when the number of iterations varies from 2 to 6.

7.2 Application to Image Encryption

Besides the speech application, the subspace-based encryption scheme is applied on image. Figure 10 shows an example of this application. Figure 10a, b, and c show the original image, the encrypted image, and the decrypted one, respectively. As one can see, there is no distinguishable difference between the original and recovered

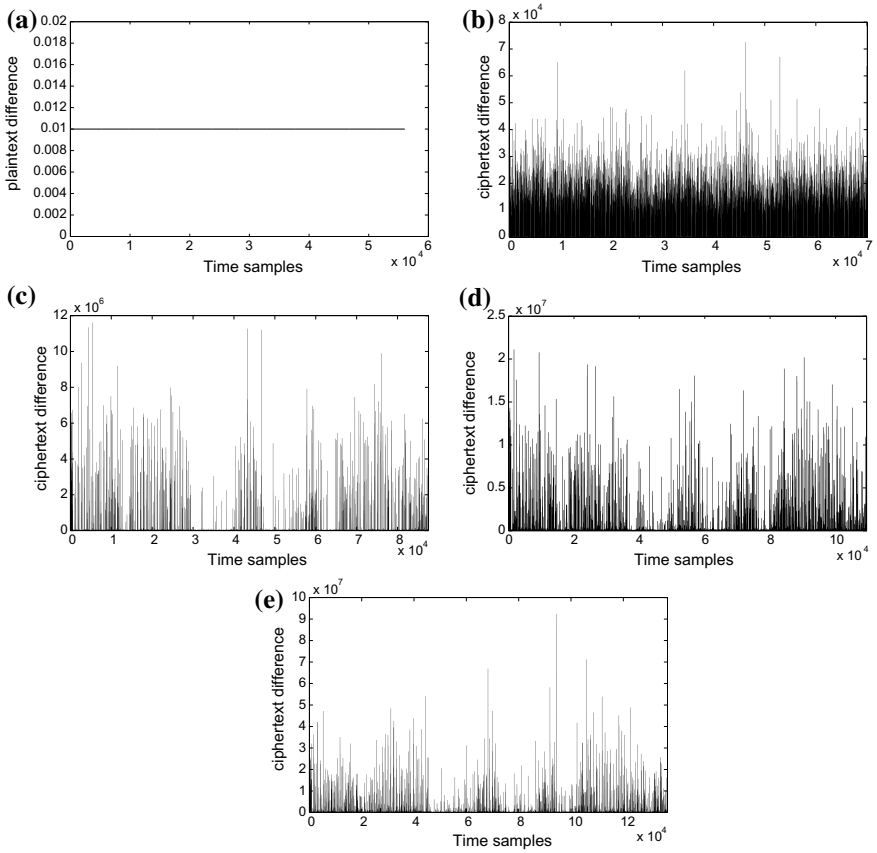


Fig. 8 A comparison in terms of sensitivity levels to a 0.01 plain-text mismatch between iterative subspace-based encryption schemes for different iterations, **a** plain-text mismatch, cipher-text difference for: **b** 1-round, **c** 2-rounds, **d** 3-rounds, **e** 4-rounds

images while the encrypted image is visually well protected. Figure 11a and b show a comparison in terms of sensitivity levels to plain-text mismatches of 0.1 and 0.01, respectively, when we apply subspace-based encryption scheme on the image used previously. One can see that the sensitivity of the subspace-based encryption scheme, revealed by the cipher-text difference level, is at lower levels when compared to the speech application's case in terms of empirical measurements. However, Fig. 12 shows, visually, the high level of sensitivity of the subspace-based encryption scheme when applied on an image. One can see that for a very small key mismatch, it is impossible to recover the original image. As it can be seen, the decrypted image looks like an encrypted one. In Fig. 13, the experimental relationship between the recovery error and the value of mismatch level ε is plotted when subspace-based scheme is used in image encryption for different iterations. The value of the factor

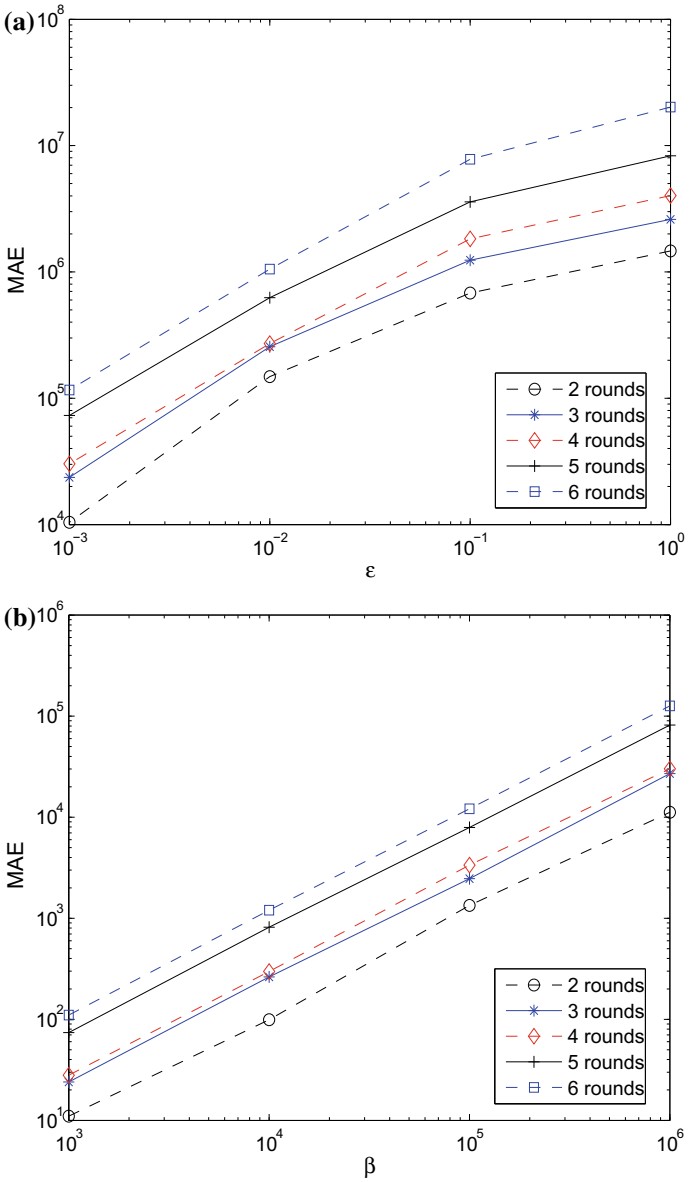


Fig. 9 The experimental relationship, in speech encryption, between the recovery error and the value of ε and β for different rounds in the iterative subspace encryption scheme, **a** $\beta = 10^6$, **b** $\varepsilon = 0.001$

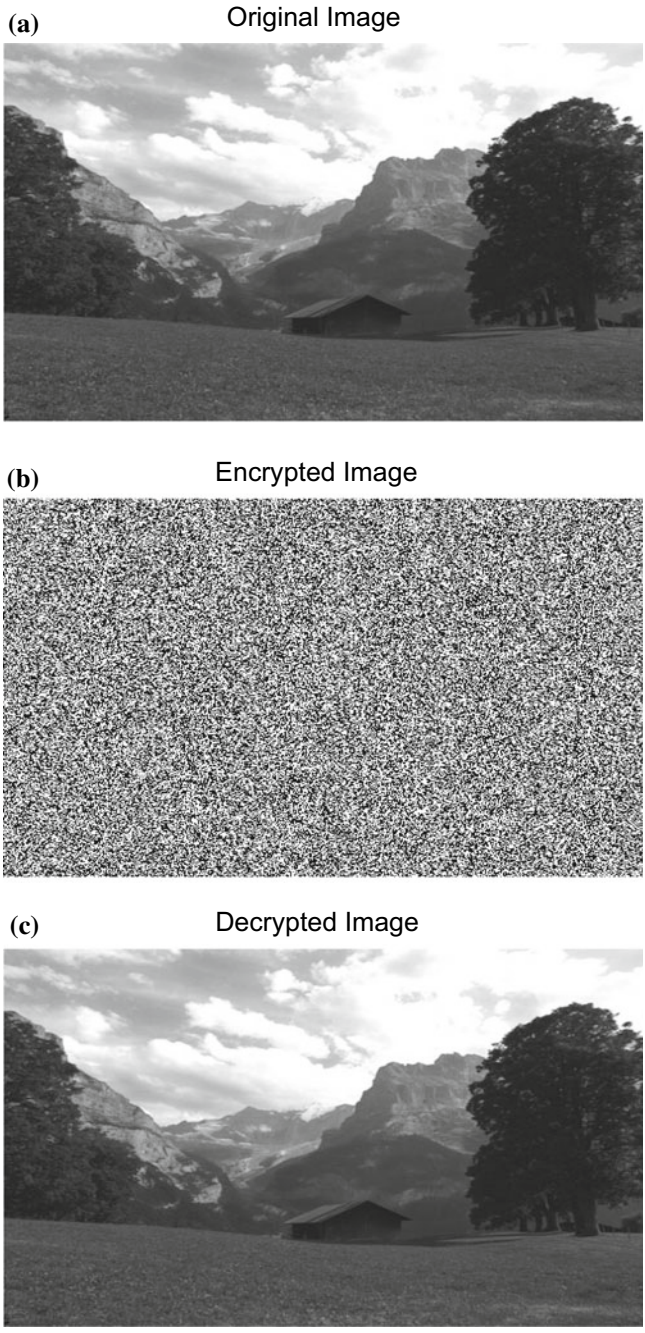


Fig. 10 An example of subspace-based image encryption, **a** Original image, **b** Encrypted image, **c** Decrypted image

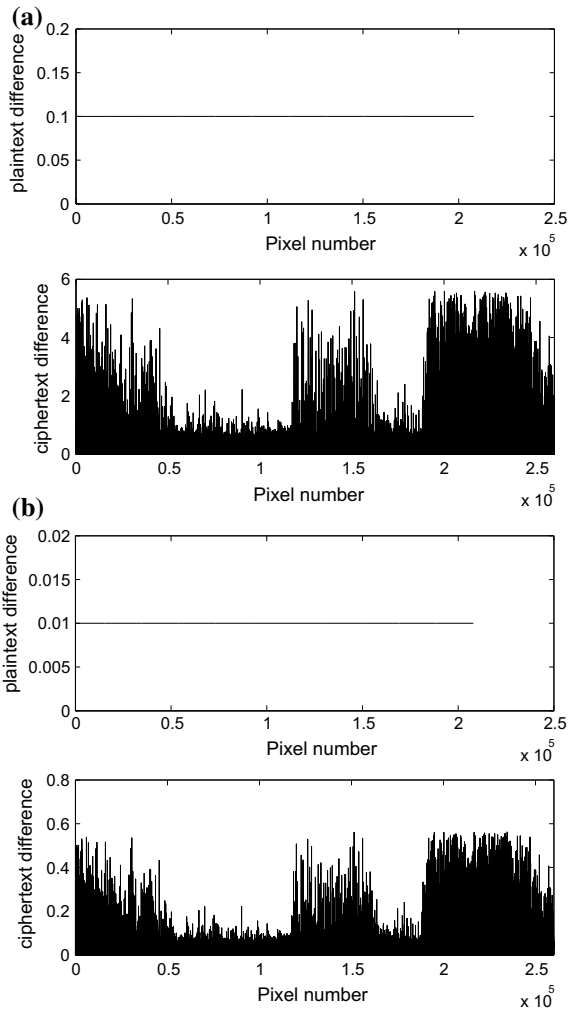


Fig. 11 Sensitivity, in image encryption, to plain-text with $\beta = 10^6$ for, **a** $\varepsilon = 0.1$, **b** $\varepsilon = 0.01$

β is equal to 10^6 and the values of ε vary from 10^{-3} to 1. For the smallest value of ε, one can see that the corresponding recovery error is high (7×10^3).

The experimental relationship, between the recovery error and the value of the factor β when the subspace-based scheme is used in image encryption, for different iterations, is shown in Fig. 14. One can see that there is a linear relationship between the factor β and the recovery error.

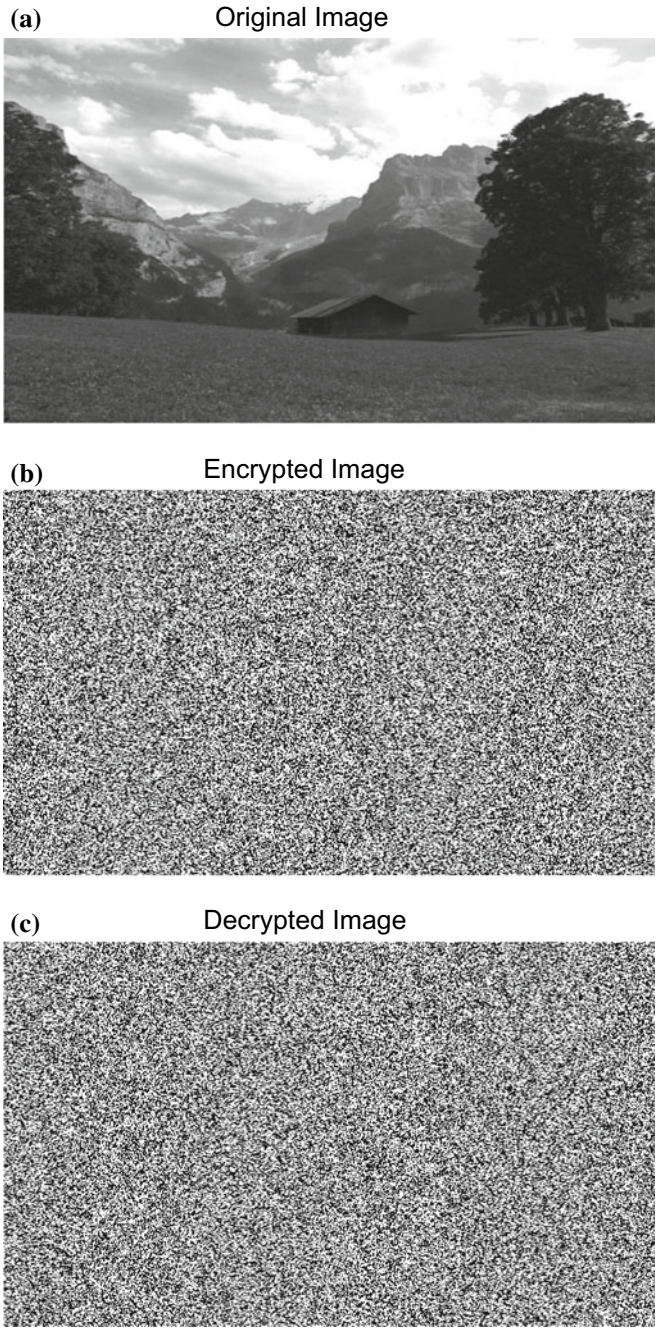


Fig. 12 An example of sensitivity of subspace-based image encryption to a very small key mismatch, **a** Original image, **b** Encrypted image, **c** Decrypted image

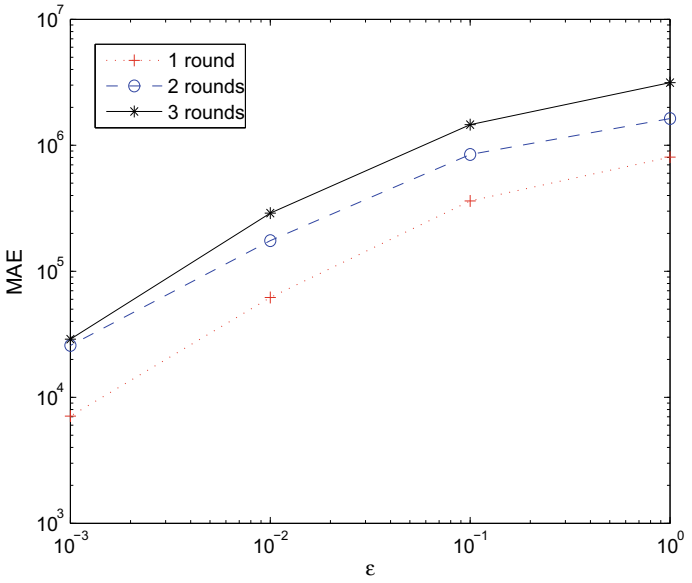


Fig. 13 The experimental relationship in subspace-based image encryption between the recovery error and the value of ε for $\beta = 10^6$

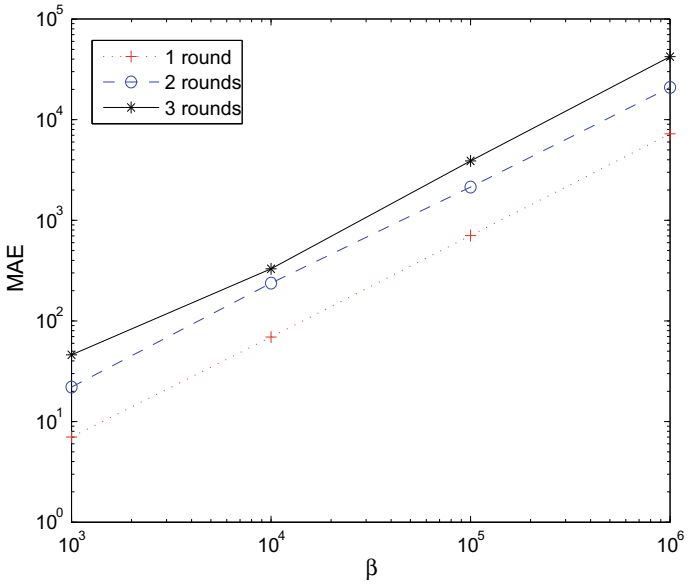


Fig. 14 The experimental relationship, in subspace-based image encryption, between the recovery error and the value of β for $\varepsilon = 0.001$

8 Conclusion

In this chapter, an investigation of the opportunity of using techniques based on the subspace concept in the encryption field is conducted. First, the investigation starts with studying the application of the Blind Source Separation (BSS) in the encryption domain. Analysis of the robustness characteristics of some BSS-based encryption techniques shows weaknesses from a cryptographic point of view. Then, a new approach based on the subspace concept is presented in order to bypass the above weaknesses. The subspace technique is applied in speech and image encryption. An iterative version of the subspace-based encryption is developed. The need for iterations is a known issue in the design of cryptographic algorithms. In the subspace-based encryption algorithm, this need is motivated by the added value, from a cryptographic robustness point of view, provided by the application of successive iterations. Of course, iterations do not have an impact on the quality of recovering the original plain-text. On another hand, several simulations and cryptanalysis tests were conducted to evaluate the robustness of the subspace-based schemes. Experimental results and discussion confirm an enhancement in security level with respect to BSS-based encryption techniques. These results show a new direction, for using nonconventional approach in encryption domain, inspired from signal processing field.

References

1. Delfs, H., Knebl, H.: Introduction to Cryptography: Principles and Applications. Springer, Berlin (2002)
2. Lin, Q.-H., Yin, F.-L.: Blind source separation applied to image cryptosystems with dual encryption. Electron. Lett. **38**(19), 1092–1094 (2002)
3. Lin, Q.-H., Yin, F.-L., Zheng, Y.-R.: Secure image communication using blind source separation. In: IEEE 6th CAS Symposium on Emerging Technologies: Mobile and Wireless CO. Shanghai, China, 31 May–2 June 2004
4. Lin, Q.-H., Yin, F.-L., Mei, T.-M., Liang, H.-L.: A speech encryption algorithm based on blind source separation. In: Proceedings of International Conference on Communication, Circuits System: Signal Processing, Circuits System, 2004, vol. II, pp. 1013–1017
5. Lin, Q.-H., Yin, F.-L., Mei, T.-M., Liang, H.: A blind source separation based method for speech encryption. IEEE Trans. Circuits Syst. I **53**(6), 1320–1328 (2006)
6. Lin, Q., Yin, F.: Image cryptosystems based on blind source separation. In: Proceedings of the 2003 International Conference on Neural Networks and Signal Processing (ICNNSP2003), vol. 2, pp. 1366–1369. IEEE (2003)
7. Lin, Q., Yin, F., Liang, H.: Blind source separation-based encryption of images and speeches. In: Advances in Neural Networks—ISNN: Proceedings, Part II, series. Lecture Notes in Computer Science, vol. 3497, pp. 544–549. Springer, Berlin, Heidelberg (2005)
8. Lin, Q.-H., Yin, F.-L., Liang, H.-L.: A fast decryption algorithm for BSS-based image encryption. In: Advances in Neural Networks—ISNN: Proceedings, Part III, series. Lecture Notes in Computer Science, vol. 3973, pp. 318–325. Springer, Berlin, Heidelberg (2006)
9. Li, S., Li, C., Lo, K.-T., Chen, G.: Cryptanalyzing of an encryption scheme based on blind source separation. IEEE Trans. Circuits Syst. I **55**(4), 1055–1063 (2008)

10. Mermoul, A., Belouchrani, A.: A subspace-based method for speech encryption. In: Proceedings of the 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA), pp. 538–541, Kuala Lumpur, Malaysia, 10–13 May 2010
11. Mermoul, A.: An iterative speech encryption scheme based on subspace technique. In: Proceedings of the 7th International Workshop on Systems, Signal Processing and their Applications (WOSSPA), pp. 361–364, Tipaza, Algeria, 09–11 May 2011
12. Anderson, R.: Security Engineering—A Guide to Building Dependable Distributed Systems. Wiley, New York (2001)
13. Mermoul, A., Absi, A.M.: On the relationship between research and community needs among the Arab-Muslim civilization heritage: an example in cryptology. In: Proceedings of the International Forum on Engineering Education. Sharjah, U.A.E, Feb 2006
14. Merayati, M., Alam, Y.M., Attayane, M.H.: Cryptography and Cryptanalysis among Arabs, pp. 39–40. Publications of the campus of the Arab language, Damascus (1987)
15. Shannon, C.E.: Communication theory of secrecy systems. Bell Syst. Tech. J. **28**, 656–715 (1949)
16. Diffie, W., Hellman, M.E.: New directions in cryptography. IEEE Trans. Inf. Theory, **IT-22**, 644–654, (1976) (Invited Paper)
17. Rivest, R., Shamir, A., Adleman, L.: A method for obtaining digital signatures and public-key cryptosystems. Commun. ACM **21**(2), 120–126, 1978
18. Mermoul, A.: Side channel attacks on cryptographic implementations. In: 1^{st} International Symposium on Electromagnetism, Satellites and Cryptography, Jijel, Algeria, 19–21 July 2005
19. Kahn, D.: The Code Breakers, 2nd edn, pp. 80–81. McGraw-Hill Inc, New York (1973)
20. Kocher, P.: Timing attacks on implementations of Diffie-Hellman, RSA, DSS, and other systems. In: Advances in Cryptology—CRYPTO'96 Proceedings, pp. 104–113. Springer (1996)
21. Kocher, P., Jaffe, J., Jun, B.: Differential power analysis: leaking secrets. In: Advances in Cryptology—Proceedings of Crypto 99. LNCS 1666, pp. 388–397. Springer
22. Gandolfi, K., Mourtel, C., Olivier, F.: Electromagnetic analysis: concrete results. In: etin Kaya Ko, David Naccache, and Christof Paar (eds.) Proceedings of Cryptographic Hardware and Embedded Systems (CHES 2001). Lecture Notes in Computer Science, vol. 2162, pp. 251–261. Springer (2001)
23. Agrawal, D., Archambeault, B., Chari, S., Rao, J.R., Rohatgi, P.: Advances in side channel cryptanalysis: electormagnetic analysis and template attacks. RSA Laboratories Cryotobytes **6**(1), 20–32 (2003)
24. NACSIM 5000: Tempest Fundamentals, National Security Agency, Fort George G.Meade, Maryland. Feb. 1982. Partially declassified also available at http://cryptome.org/nacsim-5000.htm
25. Swenson, C.: Modern Cryptanalysis: Techniques for Advanced Code Breaking. Wiley, New Jersey (2008)
26. Belouchrani, A., Cardoso, J.F.: A maximum likelihood source separation for discrete sources. Proc. EUSIPCO **2**, 768–771 (1994)
27. Taleb, A., Jutten, C.: On underdetermined source separation. In: Proceedings of IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP 99), vol. 3, pp. 1445–1448, Phoenix, Ariz, USA, March 1999
28. Aissa EL Bey, D., Abed-Meraim, K., Belouchrani, A., Grenier, Y.: Underdetermined blind separation of nondisjoint sources in the time-frequency domain. IEEE Trans. Signal Process. **55**(3), 897–907 (2007)
29. Kasprzak, W., Cichocki, A.: Hidden image separation from incomplete image mixtures by independent component analysis. In: Proceedings of the 13th International Conference on Pattern Recognition, 1996, vol. II, pp. 394–398
30. IEEE Computer Society: IEEE standard for binary floating-point arithmetic. ANSSI/IEEE Std. 754-1985, 1985
31. Schneier, B.: Applied crytpography: protocols, algorithms and source code in C, 2nd edn, p. 758. Wiley, New York (1996)

32. Mermoul, A., Belouchrani, A.: Subspace-based technique for speech encryption. Digital Signal Process. **22**(2), 298–303 (2012)
33. Gutman, P.: Cryptographic Security Architecture: Design and Verification. Springer, New York (2004)
34. Goldreich, O.: Modern Cryptography, Probalistic Proofs and Pseudo-randomness. Springer, Berlin (1999)
35. Sklar, B.: Digital Communications: Fundamentals and Applications. Prentice-Hall Intenational Inc, New Jersey (1988)
36. Soto, J.: Statistical testing of random number generators. http://csrc.nist.gov/groups/ST/toolkit/rng/documents/nissc-paper.pdf
37. Maurer, U.: A universal statistical test for random bit generators. J. Cryptol. **5**, 89–105 (1992)
38. Mirza, F.: Block ciphers and cryptanalysis. http://fmirza.seecs.nust.edu.pk (2012)
39. National Institute of Standards and Technology (US): Specification for the advanced encryption standard (AES). Federal Information Processing Standards Publication 197 (FIPS PUB 197), Nov 2001
40. Cid, C., Murphy, S., Robshaw, M.: Algebraic Aspects of the Advanced Encryption Standard. Springer, New York (2006)
41. ITU: Methods for Subjective Determination of Transmission Quality, 1996, ITU-T Rec. P.800