RF and Microwave Handbook
Second Edition

RF AND MICROWAVE APPLICATIONS AND SYSTEMS

Editor-in-Chief
Mike Golio

Managing Editor
Janet Golio

CRC Press
Taylor & Francis Group

The RF and Microwave Handbook

Second Edition

Editor-in-Chief

Mike Golio

HVVi Semiconductors, Inc.
Phoenix, Arizona, U.S.A.

Managing Editor

Janet Golio

RF and Microwave Applications and Systems

RF and Microwave Circuits, Measurements, and Modeling

RF and Microwave Passive and Active Technologies

The Electrical Engineering Handbook Series

Series Editor
Richard C. Dorf
University of California, Davis

Titles Included in the Series

The Handbook of Ad Hoc Wireless Networks, Mohammad Ilyas
The Avionics Handbook, Second Edition, Cary R. Spitzer
The Biomedical Engineering Handbook, Third Edition, Joseph D. Bronzino
The Circuits and Filters Handbook, Second Edition, Wai-Kai Chen
The Communications Handbook, Second Edition, Jerry Gibson
The Computer Engineering Handbook, Vojin G. Oklobdzija
The Control Handbook, William S. Levine
The CRC Handbook of Engineering Tables, Richard C. Dorf
The Digital Avionics Handbook, Second Edition Cary R. Spitzer
The Digital Signal Processing Handbook, Vijay K. Madisetti and Douglas Williams
The Electrical Engineering Handbook, Second Edition, Richard C. Dorf
The Electric Power Engineering Handbook, Second Edition, Leonard L. Grigsby
The Electronics Handbook, Second Edition, Jerry C. Whitaker
The Engineering Handbook, Third Edition, Richard C. Dorf
The Handbook of Formulas and Tables for Signal Processing, Alexander D. Poularikas
The Handbook of Nanoscience, Engineering, and Technology, Second Edition William A. Goddard, III, Donald W. Brenner, Sergey E. Lyshevski, and Gerald J. Iafrate
The Handbook of Optical Communication Networks, Mohammad Ilyas and Hussein T. Mouftah
The Industrial Electronics Handbook, J. David Irwin
The Measurement, Instrumentation, and Sensors Handbook, John G. Webster
The Mechanical Systems Design Handbook, Osita D.I. Nwokah and Yidirim Hurmuzlu
The Mechatronics Handbook, Second Edition, Robert H. Bishop
The Mobile Communications Handbook, Second Edition, Jerry D. Gibson
The Ocean Engineering Handbook, Ferial El-Hawary
The RF and Microwave Handbook, Second Edition, Mike Golio
The Technology Management Handbook, Richard C. Dorf
The Transforms and Applications Handbook, Second Edition, Alexander D. Poularikas
The VLSI Handbook, Second Edition, Wai-Kai Chen

RF AND MICROWAVE APPLICATIONS AND SYSTEMS

Editor-in-Chief

Mike Golio

HVVi Semiconductors, Inc.
Phoenix, Arizona, U.S.A.

Managing Editor

Janet Golio

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed in the United States of America on acid-free paper
10 9 8 7 6 5 4 3 2 1

International Standard Book Number-13: 978-0-8493-7219-3 (Hardcover)

Library of Congress Cataloging-in-Publication Data

RF and microwave applications and systems / [Ed.] Mike Golio.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-8493-7219-3 (alk. paper)
1. Microwave circuits. 2. Radio circuits. 3. Wireless communication systems. I. Golio, John Michael, 1954- II. Title.

TK7876.R486 2008
621.381'32--dc22 2007016317

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

SECTION II Systems Considerations

Preface

The first edition of the *RF and Microwave Handbook* was published in 2000. The project got off to an inauspicious start when 24 inches of snow fell in Denver the evening before the advisory board planned to hold their kick-off meeting. Two members of the board were trapped for days in the Denver airport since planes were not arriving or leaving. Because of road closures, one member was stranded only miles away from the meeting in Boulder. And the remainder of the board was stranded in a Denver hotel 10 miles from the airport. Despite this ominous beginning, a plan was formed, expert authors recruited, and the book was developed and published. The planning and development of this second edition have been very smooth and uneventful in comparison to our first efforts. Since publication in 2000, the value of the *RF and Microwave Handbook* has been recognized by thousands of engineers throughout the world. Three derivative handbooks have also been published and embraced by the microwave industry. The advisory board believes that this edition will be found to be of even greater value than the first edition.

Prior to the 1990s, microwave engineering was employed almost exclusively to address military, satellite, and avionics applications. In 1985, there were a few limited applications of RF and microwave systems that laymen might be familiar with such as satellite TV and the use of satellite communications for overseas phone calls. Pagers were also available but not common. In contrast, by 1990 the wireless revolution had begun. Cell phones were becoming common and new applications of wireless technology were emerging every day. Companies involved in wireless markets seemed to have a license to print money. At the time of the introduction of the first edition of the *RF and Microwave Handbook*, wireless electronic products were pervasive, but relatively simple, early generations of the advanced wireless products available today. At present, the number of people using wireless voice and data systems continues to grow. New systems such as 3G phones, 4G phones, and WiMAX represent emerging new wireless markets with significant growth potential. All of these wireless products are dependent on the RF and microwave component and system engineering, which is the subject of this book. During this time the military, satellite, and avionics systems have also become increasingly complex. The research and development that drives these applications continues to serve as the foundation for most of the commercial wireless products available to consumers.

This edition of the handbook covers issues of interest to engineers involved in RF/microwave system and component development. The second edition includes significantly expanded topic coverage as well as updated or new articles for most of the topics included in the first edition. The expansion of material has prompted the division of the handbook into three independent volumes of material. The chapters are aimed at working engineers, managers, and academics who have a need to understand microwave topics outside their area of expertise. Although the book is not written as a textbook, researchers and students will find it useful. Most of the chapters provide extensive references so that they will not only explain fundamentals of each field, but also serve as a starting point for further in-depth research.

This book, *RF and Microwave Applications and Systems*, includes a wide range of articles that discuss RF and microwave systems used for communications, radar and heating applications. Commercial, avionics, medical, and military applications are addressed. An overview of commercial communications systems is provided in an introductory chapter titled "Terrestrial and Satellite Mobile Radio Systems." Past, current and emerging cellular systems, navigation systems, and satellite-based systems are discussed. Specific voice and data commercial systems are investigated more thoroughly in individual chapters that follow. Detailed discussions of military electronics, avionics, and radar (both military and automotive) are provided in separate chapters. A chapter focusing on RF/microwave energy used for therapeutic medicine is also provided.

Systems considerations including thermal, mechanical, reliability, power management, and safety are discussed in separate chapters. Engineering processes are also explored in chapters about corporate initiatives, cost modeling, and design reviews. A final section of the handbook considers the underlying physics of electromagnetic propagation and interference.

Acknowledgments

Although the topics and authors for this book were identified by the editor-in-chief and the advisory board, they do not represent the bulk of the work for a project like this. A great deal of the work involves tracking down those hundreds of technical experts, gaining their commitment, keeping track of their progress, collecting their manuscripts, getting appropriate reviews/revisions, and finally transferring the documents to be published. While juggling this massive job, author inquiries ranging from, "What is the required page length?", to, "What are the acceptable formats for text and figures?", have to be answered and re-answered. Schedules are very fluid. This is the work of the Managing Editor, Janet Golio. Without her efforts there would be no second edition of this handbook.

The advisory board has facilitated the book's completion in many ways. Board members contributed to the outline of topics, identified expert authors, reviewed manuscripts, and authored several of the chapters for the book.

Hundreds of RF and microwave technology experts have produced the chapters that comprise this second edition. They have all devoted many hours of their time sharing their expertise on a wide range of topics.

I would like to sincerely thank all of those listed above. Also, it has been a great pleasure to work with Jessica Vakili, Helena Redshaw, Nora Konopka, and the publishing professionals at CRC Press.

Editors

Editor-in-Chief

Mike Golio, since receiving his PhD from North Carolina State University in 1983, has held a variety of positions in both the microwave and semiconductor industries, and within academia. As Corporate Director of Engineering at Rockwell Collins, Dr. Golio managed and directed a large research and development organization, coordinated corporate IP policy, and led committees to achieve successful corporate spin-offs.

As an individual contributor at Motorola, he was responsible for pioneering work in the area of large signal microwave device characterization and modeling. This work resulted in over 50 publications including one book and a commercially available software package. The IEEE recognized this work by making Dr. Golio a Fellow of the Institute in 1996.

He is currently RF System Technologist with HVVi Semiconductor, a start-up semiconductor company. He has contributed to all aspects of the company's funding, strategies, and technical execution.

Dr. Golio has served in a variety of professional volunteer roles for the IEEE MTT Society including: Chair of Membership Services Committee, founding Co-editor of *IEEE Microwave Magazine*, and MTT-Society distinguished lecturer. He currently serves as Editor-in-chief of *IEEE Microwave Magazine*. In 2002 he was awarded the N. Walter Cox Award for exemplary service in a spirit of selfless dedication and cooperation.

He is author of hundreds of papers, book chapters, presentations and editor of six technical books. He is inventor or co-inventor on 15 patents. In addition to his technical contributions, Dr. Golio recently published a book on retirement planning for engineers and technology professionals.

Managing Editor

Janet R. Golio is Administrative Editor of *IEEE Microwave Magazine* and webmaster of www.golio.net. Prior to that she did government, GPS, and aviation engineering at Motorola in Arizona, Rockwell Collins in Iowa, and General Dynamics in Arizona. She also helped with the early development of the personal computer at IBM in North Carolina. Golio holds one patent and has written six technical papers. She received a BS in Electrical Engineering Summa Cum Laude from North Carolina State University in 1984.

When not working, Golio actively pursues her interests in archaeology, trick roping, and country western dancing. She is the author of young adult books, *A Present from the Past* and *A Puzzle from the Past*.

Advisory Board

Peter A. Blakey

Peter A. Blakey obtained a BA in applied physics from the University of Oxford in 1972, a PhD in electronic engineering from the University of London in 1976, and an MBA from the University of Michigan in 1989. He has held several different positions in industry and academia and has worked on a wide range of RF, microwave, and Si VLSI applications. Between 1991 and 1995 he was the director of TCAD Engineering at Silvaco International. He joined the Department of Electrical Engineering at Northern Arizona University in 2002 and is presently an emeritus professor at that institution.

Nick Buris

Nick Buris received his Diploma in Electrical Engineering in 1982 from the National Technical University of Athens, Greece, and a PhD in electrical engineering from the North Carolina State University, Raleigh, North Carolina, in 1986. In 1986 he was a visiting professor at NCSU working on space reflector antennas for NASA. In 1987 he joined the faculty of the Department of Electrical and Computer Engineering at the University of Massachusetts at Amherst. His research work there spanned the areas of microwave magnetics, phased arrays printed on ferrite substrates, and broadband antennas. In the summer of 1990 he was a faculty fellow at the NASA Langley Research Center working on calibration techniques for dielectric measurements (space shuttle tiles at very high temperatures) and an ionization (plasma) sensor for an experimental reentry spacecraft.

In 1992 he joined the Applied Technology organization of Motorola's Paging Product Group and in 1995 he moved to Corporate Research to start an advanced modeling effort. While at Motorola he has worked on several projects from product design to measurement systems and the development of proprietary software tools for electromagnetic design. He currently manages the Microwave Technologies Research Lab within Motorola Labs in Schaumburg, Illinois. Recent and current activities of the group include V-band communications systems design, modeling and measurements of complex electromagnetic problems, RF propagation, Smart Antennas/MIMO, RFID systems, communications systems simulation and modeling, spectrum engineering, as well as TIA standards work on RF propagation and RF exposure.

Nick is a senior member of the IEEE, and serves on an MTT Technical Program Committee. He recently served as chair of a TIA committee on RF exposure and is currently a member of its Research Division Committee.

Lawrence P. Dunleavy

Dr. Larry Dunleavy, along with four faculty colleagues established University of South Florida's innovative Center for Wireless and Microwave Information Systems (WAMI Center—http://ee.eng.usf.edu/WAMI).

In 2001, Dr. Dunleavy co-founded Modelithics, Inc., a USF spin-off company, to provide a practical commercial outlet for developed modeling solutions and microwave measurement services (www.modelithics.com), where he is currently serving as its president.

Dr. Dunleavy received his BSEE degree from Michigan Technological University in 1982 and his MSEE and PhD in 1984 and 1988, respectively, from the University of Michigan. He has worked in industry for E-Systems (1982–1983) and Hughes Aircraft Company (1984–1990) and was a Howard Hughes Doctoral Fellow (1984–1988). In 1990 he joined the Electrical Engineering Department at the University of South Florida. He maintains a position as professor in the Department of Electrical Engineering. His research interests are related to microwave and millimeter-wave device, circuit, and system design, characterization, and modeling. In 1997–1998, Dr. Dunleavy spent a sabbatical year in the noise metrology laboratory at the National Institute of Standards and Technology (NIST) in Boulder, Colorado. Dr. Dunleavy is a senior member of IEEE and is very active in the IEEE MTT Society and the Automatic RF Techniques Group (ARFTG). He has authored or co-authored over 80 technical articles.

Jack East

Jack East received his BSE, MS, and PhD from the University of Michigan. He is presently with the Solid State Electronics Laboratory at the University of Michigan conducting research in the areas of high-speed microwave device design, fabrication, and experimental characterization of solid-state microwave devices, nonlinear and circuit modeling for communications circuits and low-energy electronics, and THz technology.

Patrick Fay

Patrick Fay is an associate professor in the Department of Electrical Engineering at the University of Notre Dame, Notre Dame, Indiana. He received his PhD in Electrical Engineering from the University of Illinois at Urbana-Champaign in 1996 after receiving a BS in Electrical Engineering from Notre Dame in 1991. Dr. Fay served as a visiting assistant professor in the Department of Electrical and Computer Engineering at the University of Illinois at Urbana-Champaign in 1996 and 1997, and joined the faculty at the University of Notre Dame in 1997.

His educational initiatives include the development of an advanced undergraduate laboratory course in microwave circuit design and characterization, and graduate courses in optoelectronic devices and electronic device characterization. He was awarded the Department of Electrical Engineering's IEEE Outstanding Teacher Award in 1998–1999. His research interests include the design, fabrication, and characterization of microwave and millimeter-wave electronic devices and circuits, as well as high-speed optoelectronic devices and optoelectronic integrated circuits for fiber optic telecommunications. His research also includes the development and use of micromachining techniques for the fabrication of microwave components and packaging. Professor Fay is a senior member of the IEEE, and has published 7 book chapters and more than 60 articles in refereed scientific journals.

David Halchin

David Halchin has worked in RF/microwaves and GaAs for over 20 years. During this period, he has worn many hats including engineering and engineering management, and he has worked in both academia and private industry. Along the way, he received his PhD in Electrical Engineering from North Carolina State University. Dave currently works for RFMD, as he has done since 1998. After a stint as a PA designer, he was moved into his current position managing a modeling organization within RFMD's Corporate Research and Development organization. His group's responsibilities include providing compact models for circuit simulation for both GaAs active devices and passives on GaAs. The group also provides compact models for a handful of Si devices, behavioral models for power amplifier assemblies, and physics-based simulation for GaAs device development. Before joining RFMD, Dave spent time at Motorola and Rockwell working in the GaAs device development and modeling areas. He is a member of the IEEE-MTT and EDS. Dave is currently a member of the executive committee for the Compound IC Symposium.

Alfy Riddle

Alfy Riddle is vice president of Engineering at Finesse. Before Finesse, Dr. Riddle was the principal at Macallan Consulting, a company he founded in 1989. He has contributed to the design and development of a wide range of products using high-speed, low-noise, and RF techniques. Dr. Riddle developed and marketed the Nodal circuit design software package that featured symbolic analysis and object-oriented techniques. He has co-authored two books and contributed chapters to several more. He is a member of the IEEE MTT Society, the Audio Engineering Society, and the Acoustical Society of America. Dr. Riddle received his PhD in Electrical Engineering in 1986 from North Carolina State University. When he is not working, he can be found on the tennis courts, hiking in the Sierras, taking pictures with an old Leica M3, or making and playing Irish flutes.

Robert J. Trew

Robert J. Trew received his PhD from the University of Michigan in 1975. He is currently the Alton and Mildred Lancaster Distinguished Professor of Electrical and Computer Engineering and Head of the ECE Department at North Carolina State University, Raleigh. He previously served as the Worcester Professor of Electrical and Computer Engineering and Head of the ECE Department of Virginia Tech, Blacksburg, Virginia, and the Dively Distinguished Professor of Engineering and Chair of the Department of Electrical Engineering and Applied Physics at Case Western Reserve University, Cleveland, Ohio. From 1997 to 2001 Dr. Trew was director of research for the U.S. Department of Defense with management responsibility for the $1.3 billion annual basic research program of the DOD. Dr. Trew was vice-chair of the U.S. government interagency group that planned and implemented the U.S. National Nanotechnology Initiative. Dr. Trew is a fellow of the IEEE, and was the 2004 president of the Microwave Theory and Techniques Society. He was editor-in-chief of the *IEEE Transactions on Microwave Theory and Techniques* from 1995 to 1997, and from 1999 to 2002 was founding co-editor-in-chief of the *IEEE Microwave Magazine.* He is currently the editor-in-chief of the *IEEE Proceedings.* Dr. Trew has twice been named an IEEE MTT Society Distinguished Microwave Lecturer. He has earned numerous honors, including a 2003 Engineering Alumni Society Merit Award in Electrical Engineering from the University of Michigan, the 2001 IEEE-USA Harry Diamond Memorial Award, the 1998 IEEE MTT Society Distinguished Educator Award, a 1992 Distinguished Scholarly Achievement Award from NCSU, and an IEEE Third Millennium Medal. Dr. Trew has authored or co-authored over 160 publications, 19 book chapters, 9 patents, and has given over 360 presentations

Contributors

Carl Andren
WLAN Consultant
Melbourne, Florida

Avram Bar-Cohen
University of Maryland
College Park, Maryland

James L. Bartlett
Rockwell Collins
Cedar Rapids, Iowa

Melvin L. Belcher, Jr.
Northrop Grumman Corporation
Reston, Virginia

Michael J. Biercuk
Booz Allen Hamilton/Defense Advanced Research Projects Agency
Arlington, Virginia

Balvinder S. Bisla
Intel Corporation
Santa Clara, California

Nicholas E. Buris
Motorola Labs
Schaumburg, Illinois

Diganta Das
CALCE-University of Maryland
College Park, Maryland

William Deal
Northrop Grumman Corporation
Redondo Beach, California

Stuart D. Edwards
Silhouette Medical
Salinas, California

John Fakatselis
WLAN Consultant
Melbourne, Florida

Leland M. Farrer
Farad Engineering
Ione, California

Patrick Fay
University of Notre Dame
Notre Dame, Indiana

Paul G. Flikkema
Northern Arizona University
Flagstaff, Arizona

Karl J. L. Geisler
General Dynamics—Advanced Information Systems
Minneapolis, Minnesota

Ian C. Gifford
Freescale Semiconductor, Inc.
Wakefield, Massachusetts

Mike Golio
HVVi Semiconductor
Phoenix, Arizona

Madhu S. Gupta
San Diego State University
San Diego, California

Ramesh K. Gupta
Mobile Satellite Ventures LP
Reston, Virginia

H. Mike Harris
Georgia Tech Research Institute
Atlanta, Georgia

Tatsuo Itoh
University of California
Los Angeles, California

William D. Jemison
Lafayette College
Easton, Pennsylvania

Dan E. Jenkins
Dell Inc.
Round Rock, Texas

Nils V. Jespersen
The Aerospace Corporation
Chantilly, Virginia

Allan D. Kraus
Allan D. Kraus Associates
Beachwood, Ohio

Andy D. Kucar
Radio4u.com
Ottawa, Ontario, Canada

Brent A. McDonald
Dell Inc.
Round Rock, Texas

Sudarshan Rao Nelatury
Penn State University, Erie
Erie, Pennsylvania

Josh T. Nessmith
Georgia Tech Research Institute
Atlanta, Georgia

Bob O'Hara
Cisco Systems
San Jose, Cafifornia

John M. Osepchuk
Full Spectrum Consulting
Concord, Massachusetts

Arthur C. Paolella
Artisan Laboratories Corporation
Chalfont, Pennsylvania

Jim Paviol
RF Consultant
Melbourne, Florida

Michael Pecht
CALCE-University of Maryland
College Park, Maryland

Ronald C. Petersen
R C Petersen Associates LLC
Bedminster, New Jersey

Al Petrick
Jones-Petrick and Associates
Orlando, Florida

Brian Petry
Member, IEEE
San Diego, California

Samuel O. Piper
Georgia Tech Research Institute
Georgia Institute of Technology
Atlanta, Georgia

Yongxi Qian
Microsemi Corporation
Los Angeles, California

Vesna Radisic
Northrop Grumman Corporation
Redondo Beach, California

Carey M. Rappaport
Northeastern University
Boston, Massachusetts

Alfy Riddle
Finesse, LLC
Santa Clara, California

Arye Rosen
Drexel University
Philadelphia, Pennsylvania

Harel D. Rosen
Onsite Neonatal Partners, Inc.
Voorhees, New Jersey

Matthew Sadiku
Prairie View A&M University
Prairie View, Texas

George K. Schoneman
Rockwell Collins
Cedar Rapids, Iowa

John F. Sevic
Maury Microwave
Ontario, California

Wayne Stark
University of Michigan
Ann Arbor, Michigan

John P. Wendler
Tyco Electronics Wireless Network Solutions
Lowell, Massachusetts

James C. Wiltse
Georgia Tech Research Institute
Georgia Institute of Technology
Atlanta, Georgia

John C. Zolper
Defense Advanced Research Projects Agency
Arlington, Virginia

Introduction to Microwaves and RF

Patrick Fay
University of Notre Dame

I.1 Introduction to Microwave and RF Engineering

Modern microwave and radio frequency (RF) engineering is an exciting and dynamic field, due in large part to the symbiosis between recent advances in modern electronic device technology and the explosion in demand for voice, data, and video communication capacity that started in the 1990s and continues through the present. Prior to this revolution in communications, microwave technology was the nearly exclusive domain of the defense industry; the recent and dramatic increase in demand for communication systems for such applications as wireless paging, mobile telephony, broadcast video, and tethered as well as untethered computer networks has revolutionized the industry. These communication systems are employed across a broad range of environments, including corporate offices, industrial and manufacturing facilities, infrastructure for municipalities, as well as private homes. The diversity of applications and operational environments has led, through the accompanying high production volumes, to tremendous advances in cost-efficient manufacturing capabilities of microwave and RF products. This in turn has lowered the implementation cost of a host of new and cost-effective wireless as well as wired RF and microwave services. Inexpensive handheld GPS navigational aids, automotive collision-avoidance radar, and widely available broadband digital service access are among these. Microwave technology is naturally suited for these emerging applications in communications and sensing, since the high operational frequencies permit both large numbers of independent channels for the wide variety of uses envisioned as well as significant available bandwidth per channel for high-speed communication.

Loosely speaking, the fields of microwave and RF engineering together encompass the design and implementation of electronic systems utilizing frequencies in the electromagnetic spectrum from approximately 300 kHz to over 100 GHz. The term “RF” engineering is typically used to refer to circuits and systems having frequencies in the range from approximately 300 kHz at the low end to between 300 MHz and 1 GHz at the upper end. The term “microwave engineering,” meanwhile, is used rather loosely to refer to design and implementation of electronic systems with operating frequencies in the range from 300 MHz to 1 GHz on the low end to upwards of 100 GHz. Figure I.1 illustrates schematically the electromagnetic spectrum from audio frequencies through cosmic rays. The RF frequency spectrum covers the medium frequency (MF), high frequency (HF), and very high frequency (VHF) bands, while the microwave portion of the

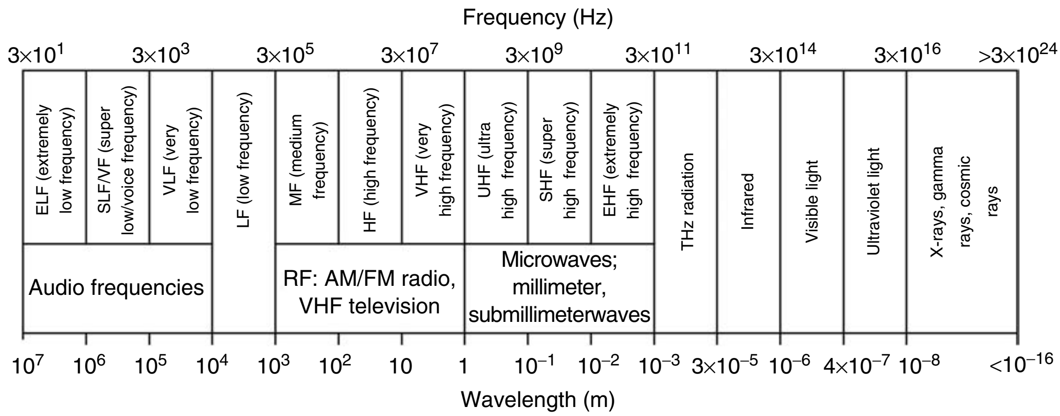


FIGURE I.1 Electromagnetic frequency spectrum and associated wavelengths.

electromagnetic spectrum extends from the upper edge of the VHF frequency range to just below the THz radiation and far-infrared optical frequencies (approximately 0.3 THz and above). The wavelength of free-space radiation for frequencies in the RF frequency range is from approximately 1 m (at 300 MHz) to 1 km (at 300 kHz), while those of the microwave range extend from 1 m to the vicinity of 1 mm (corresponding to 300 GHz) and below.

The boundary between "RF" and "microwave" design is both somewhat indistinct as well as one that is continually shifting as device technologies and design methodologies advance. This is due to implicit connotations that have come to be associated with the terms "RF" and "microwave" as the field has developed. In addition to the distinction based on the frequency ranges discussed previously, the fields of RF and microwave engineering are also often distinguished by other system features as well. For example, the particular active and passive devices used, the system applications pursued, and the design techniques and overall mindset employed all play a role in defining the fields of microwave and RF engineering. These connotations within the popular meaning of microwave and RF engineering arise fundamentally from the frequencies employed, but often not in a direct or absolute sense. For example, because advances in technology often considerably improve the high frequency performance of electronic devices, the correlation between particular types of electronic devices and particular frequency ranges is a fluid one. Similarly, new system concepts and designs are reshaping the applications landscape, with mass market designs utilizing ever higher frequencies rapidly breaking down conventional notions of microwave-frequency systems as serving "niche" markets.

The most fundamental characteristic that distinguishes RF engineering from microwave engineering is directly related to the frequency (and thus the wavelength, λ) of the electronic signals being processed. This distinction arises fundamentally from the finite speed of propagation of electromagnetic waves (and thus, by extension, currents and voltages). In free space, $\lambda = c/f$, where f is the frequency of the signal and c is the speed of light. For low-frequency and RF circuits (with a few special exceptions such as antennae), the signal wavelength is much larger than the size of the electronic system and circuit components. In contrast, for a microwave system the sizes of typical electronic components are often comparable to (i.e., within approximately 1 order of magnitude of) the signal wavelength. A schematic diagram illustrating this concept is shown in Figure I.2. As illustrated in Figure I.2, for components much smaller than the wavelength (i.e., $\ell < \lambda/10$), the finite velocity of the electromagnetic signal as it propagates through the component leads to a modest difference in phase at opposite ends of the component. For components comparable to or larger than the wavelength, however, this end-to-end phase difference becomes increasingly significant. This gives rise to a reasonable working definition of the two design areas based on the underlying approximations used in design. Since in conventional RF design, the circuit components and interconnections are generally small compared to a wavelength, they can be

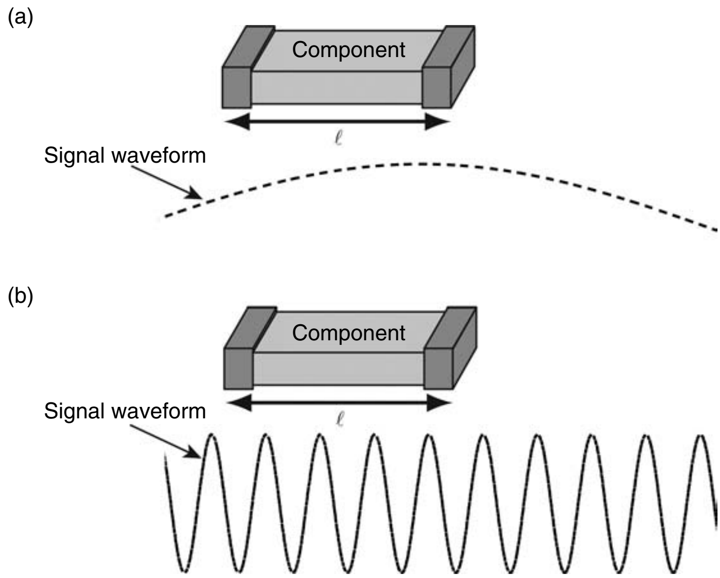


FIGURE I.2 Schematic representation of component dimensions relative to signal wavelengths. Conventional lumped-element analysis techniques are typically applicable for components for which $\ell < \lambda/10$ (a) since the phase change due to electromagnetic propagation across the component is small, while for components with $\ell > \lambda/10$ (b) the phase change is significant and a distributed circuit description is more appropriate.

modeled as lumped elements for which Kirchoff's voltage and current laws apply at every instant in time. Parasitic inductances and capacitances are incorporated to accurately model the frequency dependencies and the phase shifts, but these quantities can, to good approximation, be treated with an appropriate lumped-element equivalent circuit. In practice, a rule of thumb for the applicability of a lumped-element equivalent circuit is that the component size should be less than $\lambda/10$ at the frequency of operation. For microwave frequencies for which component size exceeds approximately $\lambda/10$, the finite propagation velocity of electromagnetic waves can no longer be as easily absorbed into simple lumped-element equivalent circuits. For these frequencies, the time delay associated with signal propagation from one end of a component to the other is an appreciable fraction of the signal period, and thus lumped-element descriptions are no longer adequate to describe the electrical behavior. A distributed-element model is required to accurately capture the electrical behavior. The time delay associated with finite wave propagation velocity that gives rise to the distributed circuit effects is a distinguishing feature of the mindset of microwave engineering.

An alternative viewpoint is based on the observation that microwave engineering lies in a "middle ground" between traditional low-frequency electronics and optics, as shown in Figure I.1. As a consequence of RF, microwaves, and optics simply being different regimes of the same electromagnetic phenomena, there is a gradual transition between these regimes. The continuity of these regimes results in constant re-evaluation of the appropriate design strategies and trade-offs as device and circuit technology advances. For example, miniaturization of active and passive components often increases the frequencies at which lumped-element circuit models are sufficiently accurate, since by reducing component dimensions the time delay for propagation through a component is proportionally reduced. As a consequence, lumped-element components at "microwave" frequencies are becoming increasingly common in systems previously based on distributed elements due to significant advances in miniaturization, even though the operational frequencies remain unchanged. Component and circuit miniaturization also leads to tighter packing of interconnects and components, potentially introducing new parasitic coupling and distributed-element effects into circuits that could previously be treated using lumped-element RF models.

The comparable scales of components and signal wavelengths has other implications for the designer as well, since neither the ray-tracing approach from optics nor the lumped-element approach from RF circuit design are valid in this middle ground. In this regard, microwave engineering can also be considered to be "applied electromagnetic engineering," as the design of guided-wave structures such as waveguides and transmission lines, transitions between different types of transmission lines, and antennae all require analysis and control of the underlying electromagnetic fields.

Guided wave structures are particularly important in microwave circuits and systems. There are many different approaches to the implementation of guided-wave structures; a sampling of the more common options are shown in Figure I.3. Figure I.3a shows a section of coaxial cable. In this common cable type, the grounded outer conductor shields the dielectric and inner conductor from external signals and also prevents the signals within the cable from radiating. The propagation in this structure is controlled by the dielectric properties, the cross-sectional geometry, and the metal conductivity. Figure I.3b shows a rectangular waveguide. In this structure, the signal propagates in the free space within the structure, while the rectangular metal structure is grounded. Despite the lack of an analog to the center conductor in the coaxial line, the structure supports traveling-wave solutions to Maxwell's equations, and thus can be used to transmit power along its length. The lack of a center conductor does prevent the structure from providing any path for dc along its length. The solution to Maxwell's equations in the rectangular waveguide also leads to multiple eigenmodes, each with its own propagation characteristics (e.g., characteristic impedance and propagation constant), and corresponding cutoff frequency. For frequencies above the cutoff frequency, the mode propagates down the waveguide with little loss, but below the cutoff frequency the mode is

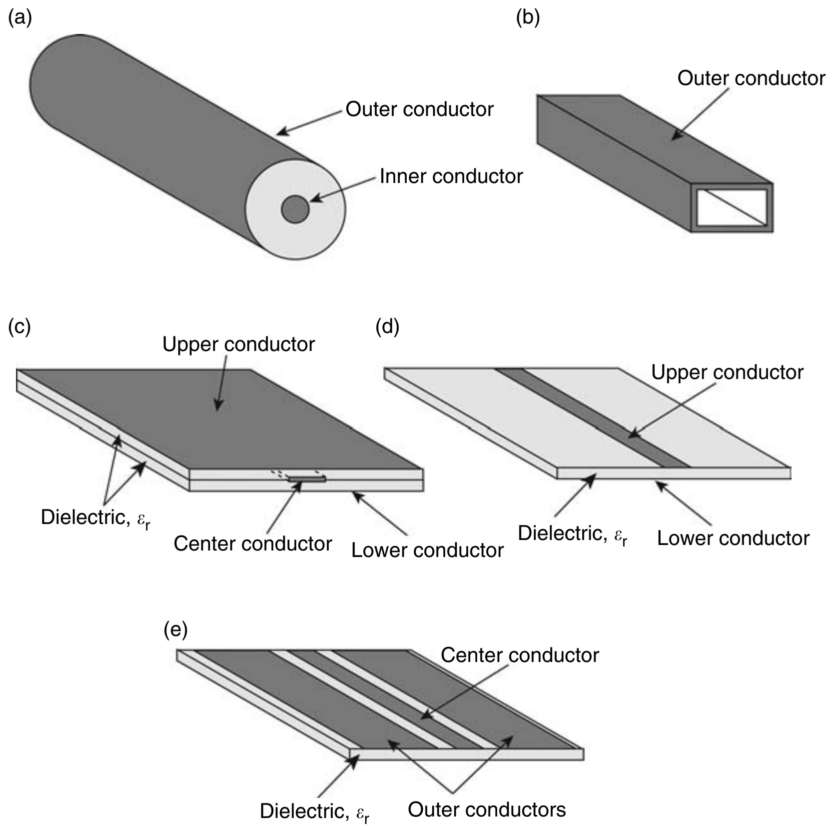


FIGURE I.3 Several common guided-wave structures. (a) coaxial cable, (b) rectangular waveguide, (c) stripline, (d) microstrip, and (e) coplanar waveguide.

evanescent and the amplitude falls off exponentially with distance. Since the characteristic impedance and propagation characteristics of each mode are quite different, in many systems the waveguides are sized to support only one propagating mode at the frequency of operation. While metallic waveguides of this type are mechanically inflexible and can be costly to manufacture, they offer extremely low loss and have excellent high-power performance. At W-band and above in particular, these structures currently offer much lower loss than coaxial cable alternatives. Figure I.3c through I.3e show several planar structures that support guided waves. Figure I.3c illustrates the stripline configuration. This structure is in some ways similar to the coaxial cable, with the center conductor of the coaxial line corresponding to the center conductor in the stripline, and the outer shield on the coaxial line corresponding to the upper and lower ground planes in the stripline. Figures I.3d and I.3e show two planar guided-wave structures often encountered in circuit-board and integrated circuit designs. Figure I.3d shows a microstrip configuration, while Figure I.3e shows a coplanar waveguide. Both of these configurations are easily realizable using conventional semiconductor and printed-circuit fabrication techniques. In the case of microstrip lines, the key design variables are the dielectric properties of the substrate, the dielectric thickness, and the width of the top conductor. For the coplanar waveguide case, the dielectric properties of the substrate, the width of the center conductor, the gap between the center and outer ground conductors, and whether or not the bottom surface of the substrate is grounded control the propagation characteristics of the lines.

For all of these guided-wave structures, an equivalent circuit consisting of the series concatenation of many stages of the form shown in Figure I.4 can be used to model the transmission line. In this equivalent circuit, the key parameters are the resistance per unit length of the line (R), the inductance per unit length (L), the parallel conductance per unit length of the dielectric (G), and the capacitance per unit length (C). Each of these parameters can be derived from the geometry and material properties of the line. Circuits of this form give rise to traveling-wave solutions of the form

$$V(z) = V_0^+ e^{-\gamma z} + V_0^- e^{\gamma z}$$

$$I(z) = \frac{V_0^+}{Z_0} e^{-\gamma z} - \frac{V_0^-}{Z_0} e^{\gamma z}$$

In these equations, the characteristic impedance of the line, which is the constant of proportionality between the current and voltage associated with a particular traveling-wave mode on the line, is given by $Z_0 = \sqrt{(R + j\omega L)/(G + j\omega C)}$. For lossless lines, $R = 0$ and $G = 0$, so that Z_0 is real; even in many practical cases the loss of the lines is small enough that the characteristic impedance can be treated as real. Similarly, the propagation constant of the line can be expressed as $\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)}$. In this expression, α characterizes the loss of the line, and β captures the wave propagation. For lossless lines, γ is pure imaginary, and thus α is zero. The design and analysis of these guided-wave structures is treated in more detail in Chapter 30.

The distinction between RF and microwave engineering is further blurred by the trend of increasing commercialization and consumerization of systems using what have been traditionally considered to be microwave frequencies. Traditional microwave engineering, with its historically military applications, has long been focused on delivering performance at any cost. As a consequence, special-purpose devices

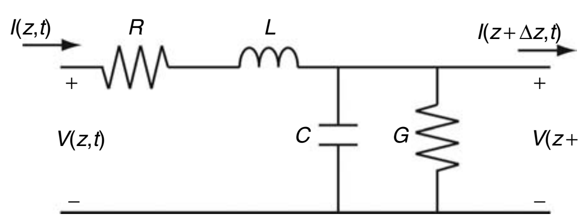


FIGURE I.4 Equivalent circuit for an incremental length of transmission line. A finite length of transmission line can be modeled as a series concatenation of sections of this form.

intended solely for use in high performance microwave systems and often with somewhat narrow ranges of applicability were developed to achieve the required performance. With continuing advances in silicon microelectronics, including Si bipolar junction transistors, SiGe heterojunction bipolar transistors (HBTs) and conventional scaled CMOS, microwave-frequency systems can now be reasonably implemented using the same devices as conventional low-frequency baseband electronics. These advanced silicon-based active devices are discussed in more detail in the companion volume *RF and Microwave Passive and Active Technologies*, Chapters 16–19. In addition, the commercialization of low-cost III–V compound semiconductor electronics, including ion-implanted metal semiconductor field-effect transistors (MESFETs), pseudomorphic and lattice-matched high electron mobility transistors (HEMTs), and III–V HBTs, has dramatically decreased the cost of including these elements in high-volume consumer systems. These compound-semiconductor devices are described in Chapters 17 and 20–22 in the *RF and Microwave Passive and Active Technologies* volume of this handbook series. This convergence, with silicon microelectronics moving ever higher in frequency into the microwave spectrum from the low-frequency side and compound semiconductors declining in price for the middle of the frequency range, blurs the distinction between "microwave" and "RF" engineering, since "microwave" functions can now be realized with "mainstream" low-cost electronics. This is accompanied by a shift from physically large, low-integration-level hybrid implementations to highly-integrated solutions based on monolithic microwave integrated circuits (MMICs) (see the companion volumes *RF and Microwave Passive and Active Technologies*, Chapters 24–25, and *RF and Microwave Circuits, Measurements, and Modeling*, Chapters 25–26). This shift has a dramatic effect not only on the design of systems and components, but also on the manufacturing technology and economics of production and implementation as well. A more complete discussion of the active device and integration technologies that make this progression possible is included in Section II of the companion volume *RF and Microwave Passive and Active Technologies* while modeling of these devices is described in Section III of *RF and Microwave Circuits, Measurements, and Modeling* in this handbook series.

Aside from these defining characteristics of RF and microwave systems, the behavior of materials is also often different at microwave frequencies than at low frequencies. In metals, the effective resistance at microwave frequencies can differ significantly from that at dc. This frequency-dependent resistance is a consequence of the skin effect, which is caused by the finite penetration depth of an electromagnetic field into conducting material. This effect is a function of frequency; the depth of penetration is given by $\delta_s = (1/\sqrt{\pi f \mu \sigma})$, where μ is the permeability, f is the frequency, and σ is the conductivity of the material. As the expression indicates, δ_s decreases with increasing frequency, and so the electromagnetic fields are confined to regions increasingly near the surface as the frequency increases. This results in the microwave currents flowing exclusively along the surface of the conductor, significantly increasing the effective resistance (and thus the loss) of metallic interconnects. Further discussion of this topic can be found in Chapter 28 of this volume and Chapter 26 of the *RF and Microwave Passive and Active Technologies* volume in this handbook series. Dielectric materials also exhibit frequency-dependent characteristics that can be important. The permeability and loss of dielectrics arises from the internal polarization and dissipation of the material. Since the polarization within a dielectric is governed by the response of the material's internal charge distribution, the frequency dependence is governed by the speed at which these charges can redistribute in response to the applied fields. For ideal materials, this dielectric relaxation leads to a frequency-dependent permittivity of the form $\varepsilon(\omega) = \varepsilon_\infty + (\varepsilon_{dc} - \varepsilon_\infty)/(1 + j\omega\tau)$, where ε_{dc} is the low-frequency permittivity, ε_∞ is the high-frequency (optical) permittivity, and τ is the dielectric relaxation time. Loss in the dielectric is incorporated in this expression through the imaginary part of ε. For many materials the dielectric relaxation time is sufficiently small that the performance of the dielectric at microwave frequencies is very similar to that at low frequencies. However, this is not universal and some care is required since some materials and devices exhibit dispersive behavior at quite low frequencies. Furthermore, this description of dielectrics is highly idealized; the frequency response of many real-world materials is much more complex than this idealized model would suggest. High-value capacitors and semiconductor devices are among the classes of devices that are particularly likely to exhibit complex dielectric responses.

In addition to material properties, some physical effects are significant at microwave frequencies that are typically negligible at lower frequencies. For example, radiation losses become increasingly important as the signal wavelengths approach the component and interconnect dimensions. For conductors and other components of comparable size to the signal wavelengths, standing waves caused by reflection of the electromagnetic waves from the boundaries of the component can greatly enhance the radiation of electromagnetic energy. These standing waves can be easily established either intentionally (in the case of antennae and resonant structures) or unintentionally (in the case of abrupt transitions, poor circuit layout, or other imperfections). Careful attention to transmission line geometry, placement relative to other components, transmission lines, and ground planes, as well as circuit packaging is essential for avoiding excessive signal attenuation and unintended coupling due to radiative effects.

A further distinction in the practice of RF and microwave engineering from conventional electronics is the methodology of testing and characterization. Due to the high frequencies involved, the impedance and standing-wave effects associated with test cables and the parasitic capacitance of conventional test probes make the use of conventional low-frequency circuit characterization techniques impractical. Although advanced measurement techniques such as electro-optic sampling can sometimes be employed to circumvent these difficulties, in general the loading effect of measurement equipment poses significant measurement challenges for debugging and analyzing circuit performance, especially for nodes at the interior of the circuit under test. In addition, for circuits employing dielectric or hollow guided-wave structures, voltage and current often cannot be uniquely defined. Even for structures in which voltage and current are well-defined, practical difficulties associated with accurately measuring such high-frequency signals make this difficult. Furthermore, since a dc-coupled time-domain measurement of a microwave signal would have an extremely wide noise bandwidth, the sensitivity of the measurement would be inadequate. For these reasons, components and low-level subsystems are characterized using specialized techniques.

One of the most common techniques for characterizing the linear behavior of microwave components is the use of *s*-parameters. While *z*-, *y*-, and *h*-parameter representations are commonly used at lower frequencies, these approaches can be problematic to implement at microwave frequencies. The use of *s*-parameters essentially captures the same information as these other parameter sets, but instead of directly measuring terminal voltages and currents, the forward and reverse traveling waves at the input and output ports are measured instead. While perhaps not intuitive at first, this approach enables accurate characterization of components at very high frequencies to be performed with comparative ease. For a two-port network, the *s*-parameters are defined by:

$$\begin{bmatrix} V_1^- \\ V_2^- \end{bmatrix} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} \begin{bmatrix} V_1^+ \\ V_2^+ \end{bmatrix}$$

where the V^- terms are the wave components traveling away from the two-port, and the V^+ terms are the incident terms. These traveling waves can be thought of as existing on "virtual" transmission lines attached to the device ports. From this definition,

$$s_{11} = \left.\frac{V_1^-}{V_1^+}\right|_{V_2^+=0}$$

$$s_{12} = \left.\frac{V_1^-}{V_2^+}\right|_{V_1^+=0}$$

$$s_{21} = \left.\frac{V_2^-}{V_1^+}\right|_{V_2^+=0}$$

$$s_{22} = \left.\frac{V_2^-}{V_2^+}\right|_{V_1^+=0}$$

To measure the s-parameters, the ratio of the forward and reverse traveling waves on the virtual input and output transmission lines is measured. To achieve the $V_1^+ = 0$ and $V_2^+ = 0$ conditions in these expressions, the ports are terminated in the characteristic impedance, Z_0, of the virtual transmission lines. Although in principle these measurements can be made using directional couplers to separate the forward and reverse traveling waves and phase-sensitive detectors, in practice modern network analyzers augment the measurement hardware with sophisticated calibration routines to remove the effects of hardware imperfections to achieve accurate *s*-parameter measurements. A more detailed discussion of *s*-parameters, as well as other approaches to device and circuit characterization, is provided in Section I in the companion volume *RF and Microwave Circuits, Measurements, and Modeling* in this handbook series.

I.2 General Applications

The field of microwave engineering is currently experiencing a radical transformation. Historically, the field has been driven by applications requiring the utmost in performance with little concern for cost or manufacturability. These systems have been primarily for military applications, where performance at nearly any cost could be justified. The current transformation of the field involves a dramatic shift from defense applications to those driven by the commercial and consumer sector, with an attendant shift in focus from design for performance to design for manufacturability. This transformation also entails a shift from small production volumes to mass production for the commercial market, and from a focus on performance without regard to cost to a focus on minimum cost while maintaining acceptable performance. For wireless applications, an additional shift from broadband systems to systems having very tightly-regulated spectral characteristics also accompanies this transformation.

For many years the driving application of microwave technology was military radar. The small wavelength of microwaves permits the realization of narrowly-focused beams to be achieved with antennae small enough to be practically steered, resulting in adequate resolution of target location. Long-distance terrestrial communications for telephony as well as satellite uplink and downlink for voice and video were among the first commercially viable applications of microwave technology. These commercial communications applications were successful because microwave-frequency carriers (f_c) offer the possibility of very wide absolute signal bandwidths (Δf) while still maintaining relatively narrow fractional bandwidths (i.e., $\Delta f / f_c$). This allows many more voice and data channels to be accommodated than would be possible with lower-frequency carriers or baseband transmission.

Among the current host of emerging applications, many are based largely on this same principle, namely, the need to transmit more and more data at high speed, and thus the need for many communication channels with wide bandwidths. Wireless communication of voice and data, both to and from individual users as well as from users and central offices in aggregate, wired communication including coaxial cable systems for video distribution and broadband digital access, fiber-optic communication systems for long- and short-haul telecommunication, and hybrid systems such as hybrid fiber-coax systems are all designed to take advantage of the wide bandwidths and consequently high data carrying capacity of microwave-frequency electronic systems. The widespread proliferation of wireless Bluetooth personal-area networks and WiFi local-area networks for transmission of voice, data, messaging and online services operating in the unlicensed ISM bands is an example of the commoditization of microwave technology for cost-sensitive consumer applications. In addition to the explosion in both diversity and capability of microwave-frequency communication systems, radar systems continue to be of importance with the emergence of nonmilitary and nonnavigational applications such as radar systems for automotive collision avoidance and weather and atmospheric sensing. Radar based noncontact fluid-level sensors are also increasingly being used in industrial process control applications. Traditional applications of microwaves in industrial material processing (primarily via nonradiative heating effects) and cooking have recently been augmented with medical uses for microwave-induced localized hyperthermia for oncological and other medical treatments.

In addition to these extensions of "traditional" microwave applications, other fields of electronics are increasingly encroaching into the microwave-frequency range. Examples include wired data networks based on coaxial cable or twisted-pair transmission lines with bit rates of over 1 Gb/s, fiber-optic communication systems with data rates well in excess of 10 Gb/s, and inexpensive personal computers and other digital systems with clock rates of well over 1 GHz. The continuing advances in the speed and capability of conventional microelectronics is pushing traditional circuit design ever further into the microwave-frequency regime. These advances have continued to push digital circuits into regimes where distributed circuit effects must be considered. While system- and board-level digital designers transitioned to the use of high-speed serial links requiring the use of distributed transmission lines in their designs some time ago, on-chip transmission lines for distribution of clock signals and the serialization of data signals for transmission over extremely high-speed serial buses are now an established feature of high-end designs within a single integrated circuit. These trends promise to both invigorate and reshape the field of microwave engineering in new and exciting ways.

I.3 Frequency Band Definitions

The field of microwave and RF engineering is driven by applications, originally for military purposes such as radar and more recently increasingly for commercial, scientific, and consumer applications. As a consequence of this increasingly diverse applications base, microwave terminology and frequency band designations are not entirely standardized, with various standards bodies, corporations, and other interested parties all contributing to the collective terminology of microwave engineering. Figure I.5 shows graphically the frequency ranges of some of the most common band designations. As can be seen from the complexity of Figure I.5, some care must be exercised in the use of the "standard" letter designations; substantial differences in the definitions of these bands exist in the literature and in practice. While the IEEE standard for radar bands [8] expressly deprecates the use of radar band designations for nonradar applications, the convenience of the band designations as technical shorthand has led to the use of these band designations in practice for a wide range of systems and technologies. This appropriation of radar band designations for other applications, as well as the definition of other letter-designated bands for other applications (e.g., electronic countermeasures) that have different frequency ranges is in part responsible for the complexity of Figure I.5. Furthermore, as progress in device and system performance opens up new system possibilities and makes ever-higher frequencies useful for new systems, the terminology of microwave engineering is continually evolving.

Figure I.5 illustrates in approximate order of increasing frequency the range of frequencies encompassed by commonly-used letter-designated bands. In Figure I.5, the dark shaded regions within the bars indicate the IEEE radar band designations, and the light cross-hatching indicates variations in the definitions by different groups and authors. The double-ended arrows appearing above some of the bands indicate other non-IEEE definitions for these letter designations that appear in the literature. For example, multiple distinct definitions of L, S, C, X, and K band are in use. The IEEE defines K band as the range from 18 to 27 GHz, while some authors define K band to span the range from 10.9 to 36 GHz, encompassing most of the IEEE's K_u, K, and K_a bands within a single band. Both of these definitions are illustrated in Figure I.5. Similarly, L band has two substantially different, overlapping definitions, with the IEEE definition of L band including frequencies from 1 to 2 GHz, with an older alternative definition of 390 MHz–1.55 GHz being found occasionally in the literature. Many other bands exhibit similar, though perhaps less extreme, variations in their definitions by various authors and standards committees. A further caution must also be taken with these letter designations, as different standards bodies and agencies do not always ensure that their letter designations are not used by others. As an example, the IEEE and U.S. military both define C, L, and K bands, but with very different frequencies; the IEEE L band resides at the low end of the microwave spectrum, while the military definition of L band is from 40 to 60 GHz. The designations (L–Y) in Figure I.5a are presently used widely in practice and the technical literature, with the newer U.S. military

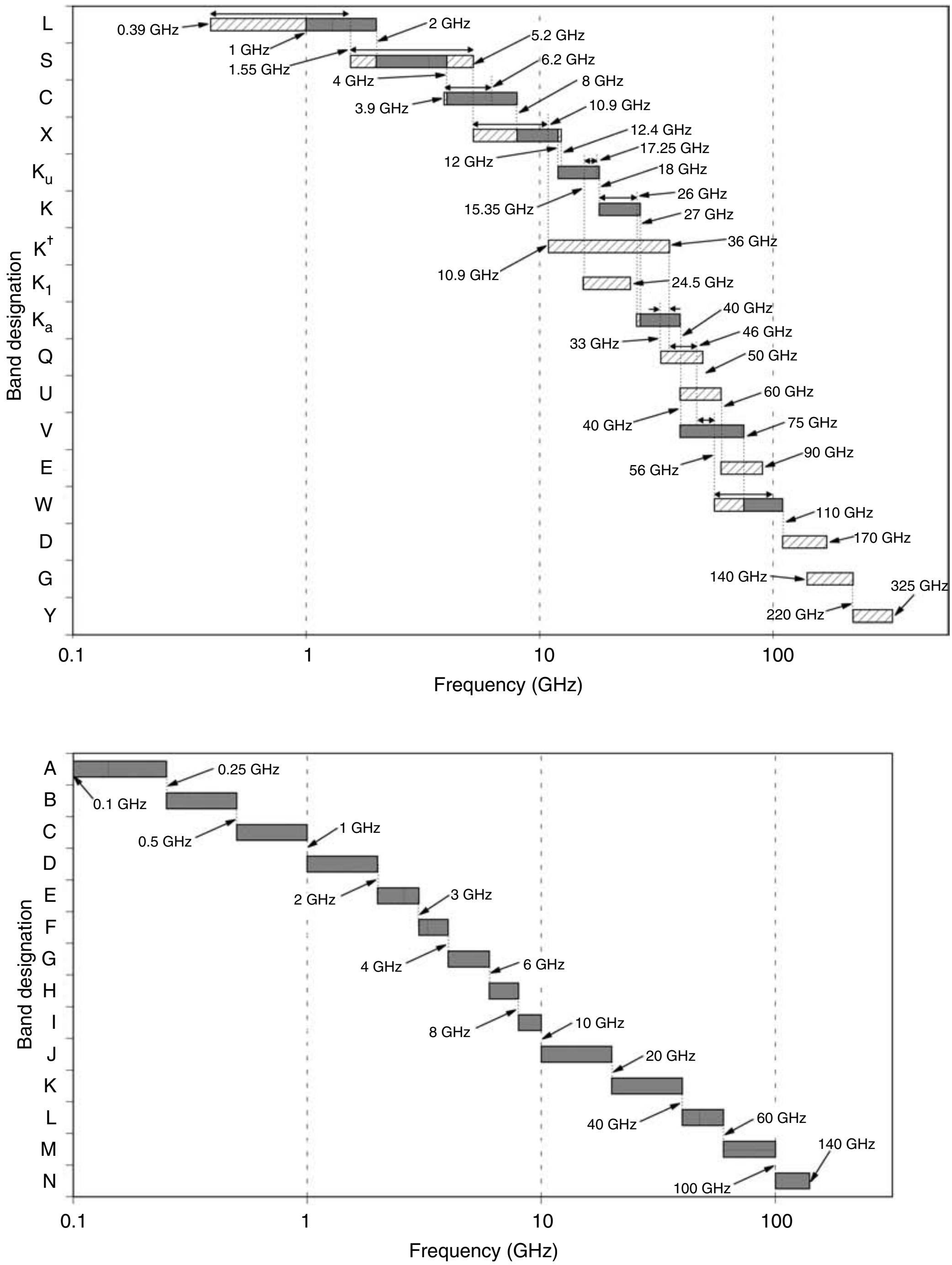


FIGURE I.5 Microwave and RF frequency band designations [1–7]. (a) Industrial and IEEE designations. Diagonal hashing indicates variation in the definitions found in literature; dark regions in the bars indicate the IEEE radar band definitions [8]. Double-ended arrows appearing above bands indicate alternative band definitions appearing in the literature, and $K^\dagger$ denotes an alternative definition for K band found in Reference [7]. (b) U.S. military frequency band designations [2–5].

designations (A–N) shown in Figure I.5b having not gained widespread popularity outside of the military community.

I.4 Overview of The RF and Microwave Handbook

The field of microwave and RF engineering is inherently interdisciplinary, spanning the fields of system architecture, design, modeling, and validation; circuit design, characterization, and verification; active and passive device design, modeling, and fabrication, including technologies as varied as semiconductor devices, solid-state passives, and vacuum electronics; electromagnetic field theory, atmospheric wave propagation, electromagnetic compatibility and interference; and manufacturing, reliability and system integration. Additional factors, including biological effects of high-frequency radiation, system cost, and market factors also play key roles in the practice of microwave and RF engineering. This extremely broad scope is further amplified by the large number of technological and market-driven design choices faced by the practitioner on a regular basis.

The full sweep of microwave and RF engineering is addressed in this three-volume handbook series. Section I of this volume focuses on system-level considerations with an application-specific focus. Typical applications, ranging from nomadic communications and cellular systems, wireless local-area networks, analog fiber-optic links, satellite communication networks, navigational aids and avionics, to radar, medical therapies, and electronic warfare applications are examined in detail. System-level considerations from the viewpoint of system integration and with focus on issues such as thermal management, cost modeling, manufacturing, and reliability are addressed in Section II of this volume in the handbook series, while the fundamental physical principles that govern the operation of devices and microwave and RF systems generally are discussed in Section III. Particular emphasis is placed on electromagnetic field theory through Maxwell's equations, free-space and guided-wave propagation, fading and multipath effects in wireless channels, and electromagnetic interference effects.

A companion volume in this handbook series, *RF and Microwave Passive and Active Technologies*, provides detailed coverage of the key active and passive device technologies encountered in microwave systems. Passive device technologies are addressed in Section I of this companion volume, including coverage of radiating elements, cables and connectors, and packaging technology, as well as in-circuit passive elements including resonators, filters, and other components. The fundamentals of active device technologies, including semiconductor diodes, transistors and integrated circuits as well as vacuum electron devices, are discussed in Section II. Key device technologies including varactor and Schottky diodes, as well as bipolar junction transistors and heterojunction bipolar transistors in both the SiGe and III-V material systems are described, as are Si MOSFETs and III-V MESFETs and HEMTs. A discussion of the fundamental physical properties at high frequencies of common materials, including metals, dielectrics, ferroelectric and piezoelectric materials, and semiconductors, is provided in Section III of this volume in the handbook series.

The third volume in this handbook series, *RF and Microwave Circuits, Measurements, and Modeling*, includes coverage of the unique difficulties and challenges encountered in accurately measuring microwave and RF devices and components. Section I of this volume discusses linear and non-linear characterization approaches, load-pull and large-signal network analysis techniques, noise measurements, fixturing and high-volume testing issues, and testing of digital systems. Consideration of key circuits for functional blocks in a wide array of system applications is addressed in Section II, including low-level circuits such as low-noise amplifiers, mixers, oscillators, power amplifiers, switches, and filters, as well as higher-level functionalities such as receivers, transmitters, and phase-locked loops. Section III of this companion volume discusses technology computer-aided design (TCAD) and nonlinear modeling of devices and circuits, along with analysis tools for systems, electromagnetics, and circuits.

References

1. Chang, K., Bahl, I., and Nair, V., *RF and Microwave Circuit and Component Design for Wireless Systems*, John Wiley & Sons, New York, 2002.
2. Collin, R. E., *Foundations for Microwave Engineering*, McGraw-Hill, New York, 1992, 2.
3. Harsany, S. C., *Principles of Microwave Technology*, Prentice Hall, Upper Saddle River, 1997, 5.
4. Laverghetta, T. S., *Modern Microwave Measurements and Techniques*, Artech House, Norwood, 1988, 479.
5. Misra, D. K., *Radio-Frequency and Microwave Communication Circuits: Analysis and Design*, John Wiley & Sons, New York, 2001.
6. Rizzi, P. A., *Microwave Engineering*, Prentice-Hall, Englewood Cliffs, 1988, 1.
7. *Reference Data for Radio Engineers*, ITT Corp., New York, 1975.
8. IEEE Std. 521-2002.

1

Overview of Microwave Engineering

Mike Golio
HVVi Semiconductor

1.1 Semiconductor Materials for RF and Microwave Applications

In addition to consideration of unique properties of metal and dielectric materials, the radio frequency (RF) and microwave engineer must also make semiconductor choices based on how existing semiconductor properties address the unique requirements of RF and microwave systems. Although semiconductor materials are exploited in virtually all electronics applications today, the unique characteristics of RF and microwave signals requires that special attention be paid to specific properties of semiconductors which are often neglected or of second-order importance for other applications. Two critical issues to RF applications are (a) the speed of electrons in the semiconductor material and (b) the breakdown field of the semiconductor material.

The first issue, speed of electrons, is clearly important because the semiconductor device must respond to high frequency changes in polarity of the signal. Improvements in efficiency and reductions in parasitic losses are realized when semiconductor materials are used which exhibit high electron mobility and velocity. Figure 1.1 presents the electron velocity of several important semiconductor materials as a

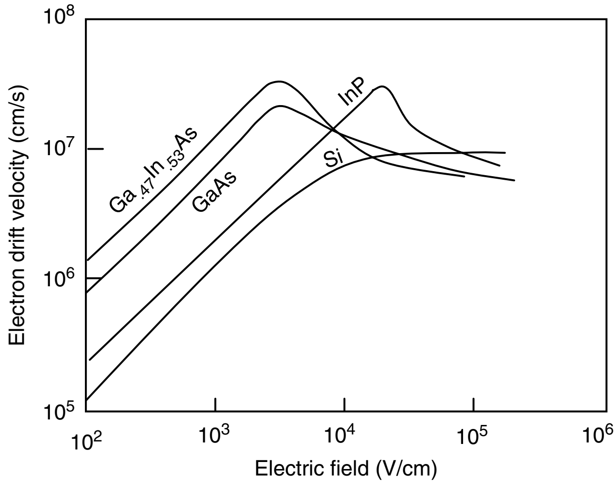


FIGURE 1.1 The electron velocity as a function of applied electric field for several semiconductor materials which are important for RF and microwave applications.

TABLE 1.1 Mobility and Breakdown Electric Field Values for Several Semiconductors Important for RF and Microwave Transmitter Applications

Property	Si	SiC	InP	GaAs	GaN
Electron mobility (cm^2/Vs)	1900	40–1000	4600	8800	1000
Breakdown field (V/cm)	3×10^5	20×10^4 to 30×10^5	5×10^5	6×10^5	$>10 \times 10^5$

function of applied electric field. The carrier mobility is given by

$$\mu_c = \frac{v}{e} \quad \text{for small values of } E \tag{1.1}$$

where v is the carrier velocity in the material and E is the electric field.

Although Silicon is the dominant semiconductor material for electronics applications today, Figure 1.1 illustrates that III–V semiconductor materials such as GaAs, GaInAs, and InP exhibit superior electron velocity and mobility characteristics relative to Silicon. Bulk mobility values for several important semiconductors are also listed in Table 1.1. As a result of the superior transport properties, transistors fabricated using III–V semiconductor materials such as GaAs, InP, and GaInAs exhibit higher efficiency and lower parasitic resistance at microwave frequencies.

From a purely technical performance perspective, the above discussion argues primarily for the use of III–V semiconductor devices in RF and microwave applications. These arguments are not complete, however. Most commercial wireless products also have requirements for high yield, high volume, low cost, and rapid product development cycles. These requirements can overwhelm the material selection process and favor mature processes and high volume experience. The silicon high volume manufacturing experience base is far greater than that of any III–V semiconductor facility.

The frequency of the application becomes a critical performance characteristic in the selection of device technology. Because of the fundamental material characteristics illustrated in Figure 1.1, Silicon device structures will always have lower theoretical maximum operation frequencies than identical III–V device structures. The higher the frequency of the application, the more likely the optimum device choice will be a III–V transistor over a Silicon transistor. Above some frequency, $f_{\text{III-V}}$, compound semiconductor devices dominate the application space, with Silicon playing no significant role in the microwave portion of the product. In contrast, below some frequency, f_{Si}, the cost and maturity advantage of Silicon provide little opportunity for III–V devices to compete. In the transition spectrum between these two frequencies Silicon and III–V devices coexist. Although Silicon devices are capable of operating above frequency f_{Si}, this operation is often gained at the expense of DC current drain. As frequency is increased above f_{Si} in the transition spectrum, efficiency advantages of GaAs and other III–V devices provide competitive opportunities for these parts. The critical frequencies, f_{Si} and f_{III-V} are not static frequency values. Rather, they are continually being moved upward by the advances of Silicon technologies—primarily by decreasing critical device dimensions.

The speed of carriers in a semiconductor transistor can also be affected by deep levels (traps) located physically either at the surface or in the bulk material. Deep levels can trap charge for times that are long compared to the signal period and thereby reduce the total RF power carrying capability of the transistor. Trapping effects result in frequency dispersion of important transistor characteristics such as transconductance and output resistance. Pulsed measurements as described in Section 1.4.4 (especially when taken over temperature extremes) can be a valuable tool to characterize deep level effects in semiconductor devices. Trapping effects are more important in compound semiconductor devices than in silicon technologies.

The second critical semiconductor issue listed in Table 1.1 is breakdown voltage. The constraints placed on the RF portion of radio electronics are fundamentally different from the constraints placed on digital circuits in the same radio. For digital applications, the presence or absence of a single electron can theoretically define a bit. Although noise floor and leakage issues make the practical limit for bit signals larger than this, the minimum amount of charge required to define a bit is very small. The bit charge minimum is also independent of the radio system architecture, the radio transmission path or the external environment. If the amount of charge utilized to define a bit within the digital chip can be reduced, then operating voltage, operating current, or both can also be reduced with no adverse consequences for the radio.

In contrast, the required propagation distance and signal environment are the primary determinants for RF signal strength. If 1 W of transmission power is required for the remote receiver to receive the signal, then reductions in RF transmitter power below this level will cause the radio to fail. Modern radio requirements often require tens, hundreds, or even thousands of Watts of transmitted power in order for the radio system to function properly. Unlike the digital situation where any discernable bit is as good as any other bit, the minimum RF transmission power must be maintained. A Watt of RF power is the product of signal current, signal voltage and efficiency, so requirements for high power result in requirements for high voltage, high current and high efficiency.

The maximum electric field before the onset of avalanche breakdown, *breakdown field*, is the fundamental semiconductor property that often limits power operation in a transistor. Table 1.1 presents breakdown voltages for several semiconductors that are commonly used in transmitter applications. In addition to Silicon, GaAs and InP, two emerging widebandgap semiconductors, SiC and GaN are included in the table. Interest from microwave engineers in these less mature semiconductors is driven almost exclusively by their attractive breakdown field capabilities. Figure 1.2 summarizes the semiconductor material application situation in terms of the power–frequency space for RF and microwave systems.

1.2 Propagation and Attenuation in the Atmosphere

Many modern RF and microwave systems are wireless. Their operation depends on transmission of signals through the atmosphere. Electromagnetic signals are attenuated by the atmosphere as they propagate from source to target. Consideration of the attenuation characteristics of the atmosphere can be critical in the design of these systems. In general, atmospheric attenuation increases with increasing frequency. As shown in Figure 1.3, however, there is significant structure in the atmospheric attenuation versus frequency plot. If only attenuation is considered, it is clear that low frequencies would be preferred for long range communications, sensor, or navigation systems in order to take advantage of the low attenuation of the atmosphere. If high data rates or large information content is required, however, higher frequencies

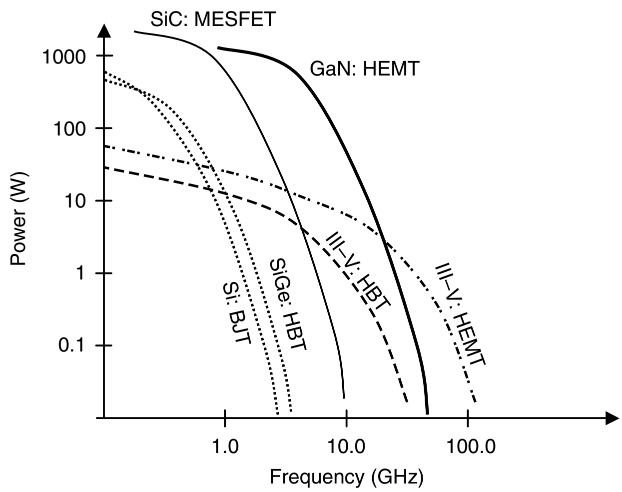


FIGURE 1.2 Semiconductor choices for RF applications are a strong function of the power and frequency required for the wireless application.

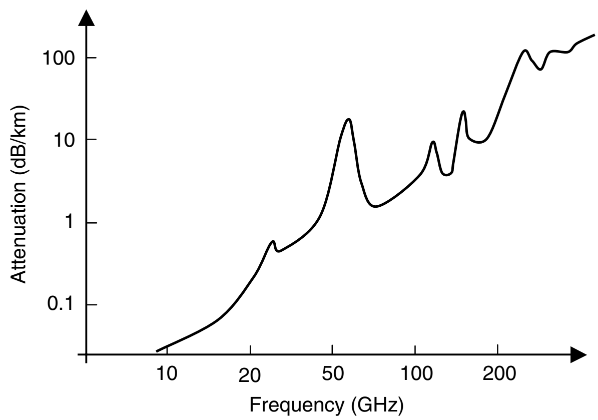


FIGURE 1.3 Attenuation of electromagnetic signals in the atmosphere as a function of frequency.

are needed. In addition to the atmospheric attenuation, the wavelengths of microwave systems are small enough to become effected by water vapor and rain. Above 10 GHz these effects become important. Above 25 GHz, the effect of individual gas molecules becomes important. Water and oxygen are the most important gases. These have resonant absorption lines at ~23, ~69, and ~120 GHz. In addition to absorption lines, the atmosphere also exhibits "windows" that may be used for communication, notably at ~38 and ~98 GHz.

RF and microwave signal propagation is also affected by objects such as trees, buildings, towers, and vehicles in the path of the wave. Indoor systems are affected by walls, doors, furniture, and people. As a result of the interaction of electromagnetic signals with objects, the propagation channel for wireless communication systems consists of multiple paths between the transmitter and receiver. Each path will experience different attenuation and delay. Some transmitted signals may experience a deep fade (large attenuation) due to destructive multipath cancellation. Similarly, constructive multipath addition can produce signals of large amplitude. Shadowing can occur when buildings or other objects obstruct the line-of-site path between transmitter and receiver.

The design of wireless systems must consider the interaction of specific frequencies of RF and microwave signals with the atmosphere and with objects in the signal channel that can cause multipath effects.

1.3 Systems Applications

There are four important classes of applications for microwave and RF systems: communications, navigation, sensors, and heating. Each of these classes of applications benefits from some of the unique properties of high-frequency electromagnetic fields.

1.3.1 Communications

Wireless communications applications have exploded in popularity over the past decade. Pagers, cellular phones, radio navigation, and wireless data networks are among the RF products that consumers are likely to be familiar with. Prior to the growth of commercial wireless communications, RF and microwave radios were in common usage for communications satellites, commercial avionics communications, and many government and military radios. All of these systems benefit from the high frequencies that offer greater bandwidth than low frequency systems, while still propagating with relatively low atmospheric losses compared to higher frequency systems.

Cellular phones are among the most common consumer radios in use today. Analog cellular (first generation or 1G cellular) operates at 900 MHz bands and was first introduced in 1983. Second generation (2G) cellular using TDMA, GSM TDMA, and CDMA digital modulation schemes came into use more than 10 years later. The 2G systems were designed to get greater use of the 1.9 GHz frequency bands than their analog predecessors. Emergence of 2.5G and 3G systems operating in broader bands as high as 2.1 GHz is occurring today. These systems make use of digital modulation schemes adapted from 2G GSM and CDMA systems. With each advance in cellular phones, requirements on the microwave circuitry have increased. Requirements for broader bandwidths, higher efficiency and greater linearity have been coupled with demands for lower cost, lighter, smaller products, and increasing functionality. The microwave receivers and transmitters designed for portable cellular phones represent one of the highest volume manufacturing requirements of any microwave radio. Fabrication of popular cell phones has placed an emphasis on manufacturability and yield for microwave radios that was unheard of prior to the growth in popularity of these products.

Other microwave-based consumer products that are growing dramatically in popularity are the wireless local area network (WLAN) or Wi-Fi and the longer range WiMAX systems. These systems offer data rates more than five times higher than cellular-based products using bandwidth at 2.4, 3.5, and 5 GHz. Although the volume demands for Wi-Fi and WiMAX components are not as high as for cellular phones, the emphasis on cost and manufacturability is still critical to these products.

Commercial communications satellite systems represent a microwave communications product that is less conspicuous to the consumer, but continues to experience increasing demand. Although the percentage of voice traffic carried via satellite systems is rapidly declining with the advent of undersea fiber-optic cables, new video and data services are being added over existing voice services. Today satellites provide worldwide TV channels, global messaging services, positioning information, communications from ships and aircraft, communications to remote areas, and high-speed data services including internet access. Allocated satellite communication frequency bands include spectrum from as low as 2.5 GHz to almost 50 GHz. These allocations cover extremely broad bandwidths compared to many other communications systems. Future allocation will include even higher frequency bands. In addition to the bandwidth and frequency challenges, microwave components for satellite communications are faced with reliability requirements that are far more severe than any earth-based systems.

Avionics applications include subsystems that perform communications, navigation, and sensor applications. Avionics products typically require functional integrity and reliability that are orders of magnitude more stringent than most commercial wireless applications. The rigor of these requirements is matched or exceeded only by the requirements for space and/or certain military applications. Avionics must function in environments that are more severe than most other wireless applications as well. Quantities of products required for this market are typically very low when compared to commercial wireless applications, for example, the number of cell phones manufactured every single working day far exceeds the number of aircraft that are manufactured in the world in a year. Wireless systems for avionics applications cover an extremely wide range of frequencies, function, modulation type, bandwidth, and power. Due to the number of systems aboard a typical aircraft, Electromagnetic Interference (EMI) and Electromagnetic Compatibility (EMC) between systems is a major concern, and EMI/EMC design and testing is a major factor in the flight certification testing of these systems. RF and microwave communications systems for avionics applications include several distinct bands between 2 and 400 MHz and output power requirements as high as 100 Watts.

In addition to commercial communications systems, military communication is an extremely important application of microwave technology. Technical specifications for military radios are often extremely demanding. Much of the technology developed and exploited by existing commercial communications systems today was first demonstrated for military applications. The requirements for military radio applications are varied but will cover broader bandwidths, higher power, more linearity, and greater levels of integration than most of their commercial counterparts. In addition, reliability requirements for these systems are stringent. Volume manufacturing levels, of course, tend to be much lower than commercial systems.

1.3.2 Navigation

Electronic navigation systems represent a unique application of microwave systems. In this application, data transfer takes place between a satellite (or fixed basestation) and a portable radio on earth. The consumer portable product consists of only a receiver portion of a radio. No data or voice signal is transmitted by the portable navigation unit. In this respect, electronic navigation systems resemble a portable paging system more closely than they resemble a cellular phone system. The most widespread electronic navigation system is GPS. The nominal GPS constellation is composed of 24 satellites in six orbital planes, (four satellites in each plane). The satellites operate in circular 20,200 km altitude (26,570 km radius) orbits at an inclination angle of 55°. Each satellite transmits a navigation message containing its orbital elements, clock behavior, system time, and status messages. The data transmitted by the satellite are sent in two frequency bands at 1.2 and 1.6 GHz. The portable terrestrial units receive these messages from multiple satellites and calculate the location of the unit on the earth. In addition to GPS, other navigation systems in common usage include NAVSTAR, GLONASS, and LORAN.

1.3.3 Sensors (Radar)

Microwave sensor applications are addressed primarily with various forms of radar. Radar is used by police forces to establish the speed of passing automobiles, by automobiles to establish vehicle speed and danger of collision, by air traffic control systems to establish the locations of approaching aircraft, by aircraft to establish ground speed, altitude, other aircraft and turbulent weather, and by the military to establish a multitude of different types of targets.

The receiving portion of a radar unit is similar to other radios. It is designed to receive a specific signal and analyze it to obtain desired information. The radar unit differs from other radios, however, in that the signal that is received is typically transmitted by the same unit. By understanding the form of the transmitted signal, the propagation characteristics of the propagation medium, and the form of the received (reflected) signal, various characteristics of the radar target can be determined including size, speed, and distance from the radar unit. As in the case of communications systems, radar applications benefit from the propagation characteristics of RF and microwave frequencies in the atmosphere. The best frequency to use for a radar unit depends upon its application. Like most other radio design decisions, the choice of frequency usually involves trade-offs among several factors including physical size, transmitted power, and atmospheric attenuation.

The dimensions of radio components used to generate RF power and the size of the antenna required to direct the transmitted signal are, in general, proportional to wavelength. At lower frequencies where wavelengths are longer, the antennae and radio components tend to be large and heavy. At the higher frequencies where the wavelengths are shorter, radar units can be smaller and lighter.

Frequency selection can indirectly influence the radar power level because of its impact on radio size. Design of high power transmitters requires that significant attention be paid to the management of electric field levels and thermal dissipation. Such management tasks are made more complex when space is limited. Since radio component size tends to be inversely proportional to frequency, manageable power levels are reduced as frequency is increased.

As in the case of all wireless systems, atmospheric attenuation can reduce the total range of the system. Radar systems designed to work above about 10 GHz must consider the atmospheric loss at the specific frequency being used in the design.

Automotive radar represents a large class of radars that are used within an automobile. Applications include speed measurement, adaptive cruise control, obstacle detection, and collision avoidance. Various radar systems have been developed for forward-, rear-, and side-looking applications.

V-band frequencies are exploited for forward looking radars. Within V-band, different frequencies have been used in the past decade, including 77 GHz for U.S. and European systems, and 60 GHz in some Japanese systems. The choice of V-band for this application is dictated by the resolution requirement, antenna size requirement and the desire for atmospheric attenuation to insure the radar is short range. The frequency requirement of this application has contributed to a slow emergence of this product into mainstream use, but the potential of this product to have a significant impact on highway safety continues to keep automotive radar efforts active.

As in the case of communications systems, avionics and military users also have significant radar applications. Radar is used to detect aircraft both from the earth and from other aircraft. It is also used to determine ground speed, establish altitude, and detect weather turbulence.

1.3.4 Heating

The most common heating application for microwave signals is the microwave oven. These consumer products operate at a frequency that corresponds to a resonant frequency of water. When exposed to electromagnetic energy at this frequency, all water molecules begin to spin or oscillate at that frequency. Since all foods contain high percentages of water, electromagnetic energy at this resonant frequency interacts with all foods. The energy absorbed by these rotating molecules is transferred to the food in the form of heat.

RF heating can also be important for medical applications. Certain kinds of tumors can be detected by the lack of electromagnetic activity associated with them and some kinds of tumors can be treated by heating them using electromagnetic stimulation.

The use of RF/microwaves in medicine has increased dramatically in recent years. RF and microwave therapies for cancer in humans are presently used in many cancer centers. RF treatments for heartbeat irregularities are currently employed by major hospitals. RF/microwaves are also used in human subjects for the treatment of certain types of benign prostrate conditions. Several centers in the United States have been utilizing RF to treat upper airway obstruction and alleviate sleep apnea. New treatments such as microwave aided liposuction, tissue joining in conjunction with microwave irradiation in future endoscopic surgery, enhancement of drug absorption, and microwave septic wound treatment are continually being researched.

1.4 Measurements

The RF/microwave engineer faces unique measurement challenges. At high frequencies, voltages and currents vary too rapidly for conventional electronic measurement equipment to gauge. Conventional curve tracers and oscilloscopes are of limited value when microwave component measurements are needed. In addition, calibration of conventional characterization equipment typically requires the use of open and short circuit standards that are not useful to the microwave engineer. For these reasons, most commonly exploited microwave measurements focus on the measurement of power and phase in the frequency domain as opposed to voltages and currents in the time domain.

1.4.1 Small Signal

Characterization of the linear performance of microwave devices, components and boards is critical to the development of models used in the design of the next higher level of microwave subsystem. At lower frequencies, direct measurement of y-, z-, or h-parameters is useful to accomplish linear characterization. As discussed in Chapter 1, however, RF and microwave design utilizes s-parameters for this application. Other small signal characteristics of interest in microwave design include impedance, VSWR, gain, and attenuation. Each of these quantities can be computed from two-port s-parameter data.

The s-parameters defined in Chapter 1 are complex quantities normally expressed as magnitude and phase. Notice that S_{11} and S_{22} can be thought of as complex reflection ratios since they represent the magnitude and phase of waves reflected from port 1 (input) and 2 (output), respectively. It is common to measure the quality of the match between components using the *reflection coefficient* defined as

$$\Gamma = |S_{11}| \tag{1.2}$$

for the input reflection coefficient of a two-port network, or

$$\Gamma = |S_{22}| \tag{1.3}$$

for the output reflection coefficient.

Reflection coefficient measurements are often expressed in dB and referred to as *return loss* evaluated as

$$L_{\text{return}} = -20\log(\Gamma). \tag{1.4}$$

Analogous to the reflection coefficient, both a forward and reverse *transmission coefficient* can be measured. The forward transmission coefficient is given as

$$T = |S_{21}| \tag{1.5}$$

while the reverse transmission coefficient is expressed

$$T = |S_{12}|\,. \tag{1.6}$$

As in the case of reflection coefficient, transmission coefficients are often expressed in dB and referred to as *gain* given by

$$G = 20\log(T). \tag{1.7}$$

Another commonly measured and calculated parameter is the *standing wave ratio* or the *voltage standing wave ratio* (VSWR). This quantity is the ratio of maximum to minimum voltage at a given port. It is commonly expressed in terms of reflection coefficient as

$$\text{VSWR} = \frac{1+\Gamma}{1-\Gamma}. \tag{1.8}$$

The vector network analyzer (VNA) is the instrument of choice for small signal characterization of high-frequency components. Figure 1.4 illustrates a one-port VNA measurement. These measurements

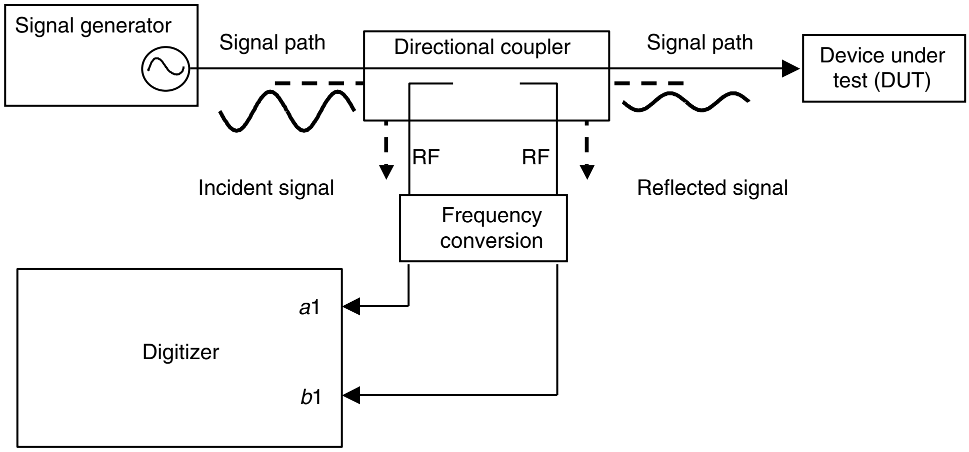


FIGURE 1.4 Vector network analyzer measurement configuration to determine *s*-parameters of a high-frequency device, component, or subsystem.

use a source with well-defined impedance equal to the system impedance and all ports of the device under test (DUT) are terminated with the same impedance. This termination eliminates unwanted signal reflections during the measurement. The port being measured is terminated in the test channel of the network analyzer that has input impedance equal to the system characteristic impedance. Measurement of system parameters with all ports terminated minimizes the problems caused by short-, open-, and test-circuit parasitics that cause considerable difficulty in the measurement of *y*- and *h*-parameters at very high frequencies. If desired, *s*-parameters can be converted to *y*- and *h*-parameters using analytical mathematical expressions.

The directional coupler shown in Figure 1.4 is a device for measuring the forward and reflected waves on a transmission line. During the network analyzer measurement, a signal is driven through the directional coupler to one port of the DUT. Part of the incident signal is sampled by the directional coupler. On arrival at the DUT port being measured, some of the incident signal will be reflected. This reflection is again sampled by the directional coupler. The sampled incident and reflected signals are then downconverted in frequency and digitized. The measurement configuration of Figure 1.4 shows only one-half of the equipment required to make full two-port *s*-parameter measurements. The *s*-parameters as defined in Chapter 1 are determined by analyzing the ratios of the digitized signal data.

For many applications, knowledge of the magnitude of the incident and reflected signals is sufficient (i.e., Γ is all that is needed). In these cases, the scalar network analyzer can be utilized in place of the VNA. The cost of the scalar network analyzer equipment is much less than VNA equipment and the calibration required for making accurate measurements is easier when phase information is not required. The scalar network analyzer measures reflection coefficient as defined in Equations 2.1 and 2.2.

1.4.2 Large Signal

Virtually all physical systems exhibit some form of nonlinear behavior and microwave systems are no exception. Although powerful techniques and elaborate tools have been developed to characterize and analyze linear RF and microwave circuits, it is often the nonlinear characteristics that dominate microwave engineering efforts. Nonlinear effects are not all undesirable. Frequency conversion circuitry, for example, exploits nonlinearities in order to translate signals from one frequency to another. Nonlinear performance characteristics of interest in microwave design include harmonic distortion, gain compression, intermodulation distortion (IMD), phase distortion, and adjacent channel power. Numerous other nonlinear phenomena and nonlinear figures-of-merit are less commonly addressed, but can be important for some microwave systems.

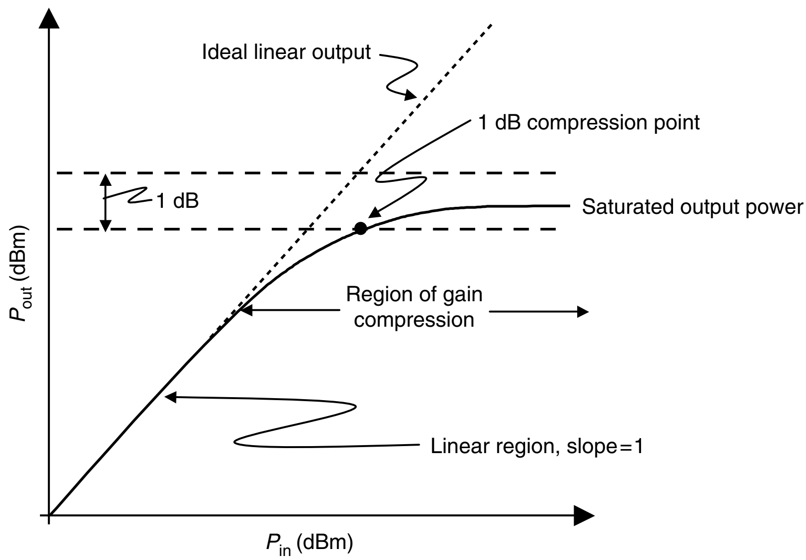


FIGURE 1.5 Output power versus input power at the fundamental frequency for a nonlinear circuit.

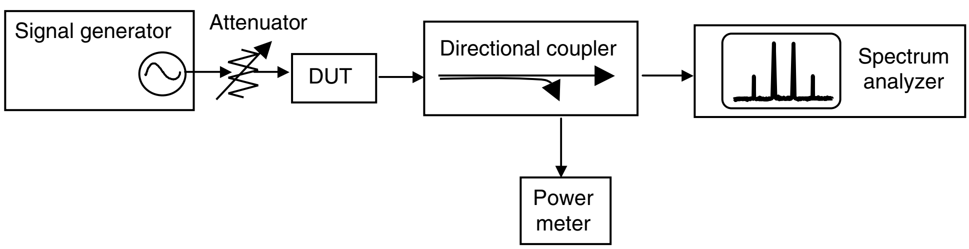


FIGURE 1.6 Measurement configuration to characterize gain compression and harmonic distortion. By replacing the signal generator with two combined signals at slightly offset frequencies, the configuration can also be used to measure intermodulation distortion.

1.4.2.1 Gain Compression

Figure 1.5 illustrates gain compression characteristics of a typical microwave amplifier with a plot of output power as a function of input power. At low power levels, a single frequency signal is increased in power level by the small signal gain of the amplifier ($P_{out} = G * P_{in}$). At lower power levels, this produces a linear P_{out} versus P_{in} plot with slope = 1 when the powers are plotted in dB units as shown in Figure 1.5. At higher power levels, nonlinearities in the amplifier begin to generate some power in the harmonics of the single frequency input signal and to compress the output signal. The result is decreased gain at higher power levels. This reduction in gain is referred to as *gain compression*. Gain compression is often characterized in terms of the power level when the large signal gain is 1 dB less than the small signal gain. The power level when this occurs is termed the *1*dB *compression point* and is also illustrated in Figure 1.5.

The microwave spectrum analyzer is the workhorse instrument of nonlinear microwave measurements. The instrument measures and displays power as a function of swept frequency. Combined with a variable power level signal source (or multiple combined or modulated sources), many nonlinear characteristics can be measured using the spectrum analyzer in the configuration illustrated in Figure 1.6.

1.4.2.2 Harmonic Distortion

A fundamental result of nonlinear distortion in microwave devices is that power levels are produced at frequencies which are integral multiples of the applied signal frequency. These other frequency components are termed *harmonics* of the fundamental signal. Harmonic signal levels are usually specified and measured relative to the fundamental signal level. The harmonic level is expressed in dBc, which designates dB relative to the fundamental power level. Microwave system requirements often place a maximum acceptable level for individual harmonics. Typically third and second harmonic levels are critical, but higher-order harmonics can also be important for many applications. The measurement configuration illustrated in Figure 1.6 can be used to directly measure harmonic distortion of a microwave device.

1.4.2.3 Intermodulation Distortion

When a microwave signal is composed of power at multiple frequencies, a nonlinear circuit will produce IMD. The IMD characteristics of a microwave device are important because they can create unwanted interference in adjacent channels of a radio or radar system. The intermodulation products of two signals produce distortion signals not only at the harmonic frequencies of the two signals, but also at the sum and difference frequencies of all of the signal's harmonics. If the two signal frequencies are closely spaced at frequencies f_c and f_m, then the IMD products located at frequencies $2f_c - f_m$ and $2f_m - f_c$ will be located very close to the desired signals. This situation is illustrated in the signal spectrum of Figure 1.7. The IMD products at $2f_c - f_m$ and $2f_m - f_c$ are third-order products of the desired signals, but are located so closely to f_c and f_m that filtering them out of the overall signal is difficult.

The spectrum of Figure 1.7 represents the nonlinear characteristics at a single power level. As power is increased and the device enters gain compression, however, harmonic power levels will grow more quickly than fundamental power levels. In general, the *n*th-order harmonic power level will increase at n times the fundamental. This is illustrated in the P_{out} versus P_{in} plot of Figure 1.8 where both the fundamental and the third-order products are plotted. As in the case of the fundamental power, third-order IMD levels will compress at higher power levels. IMD is often characterized and specified in terms of the *third-order intercept point*, IP3. This point is the power level where the slope of the small signal gain and the slope of the low power level third-order product characteristics cross as shown in Figure 1.8.

1.4.2.4 Phase Distortion

Reactive elements in a microwave system give rise to time delays that are nonlinear. Such delays are referred to as *memory effects* and result in *AM–PM distortion* in a modulated signal. AM–PM distortion creates

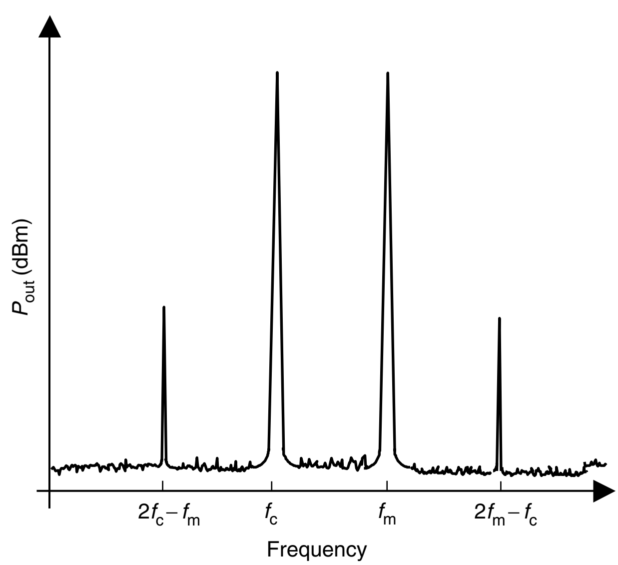


FIGURE 1.7 An illustration of signal spectrum due to intermodulation distortion from two signals at frequencies f_c and f_m.

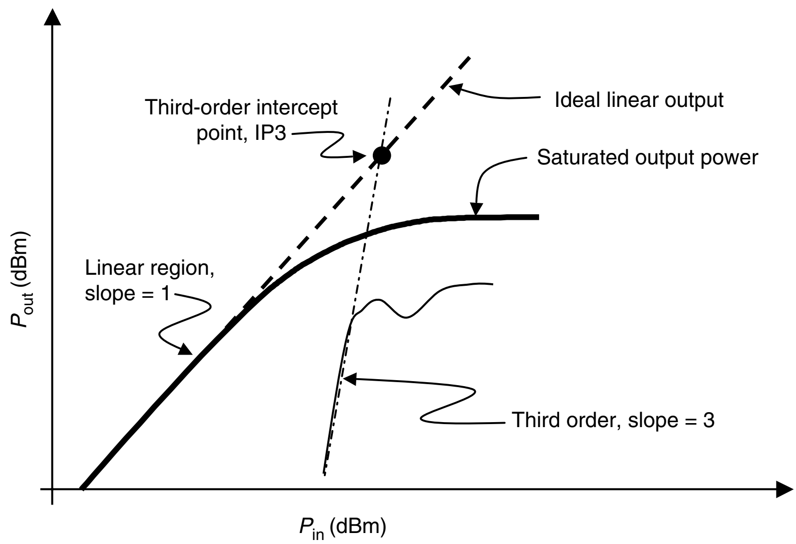


FIGURE 1.8 Relationship between signal output power and intermodulation distortion product levels.

sidebands at harmonics of a modulating signal. These sidebands are similar to the IMD sidebands, but are repeated for multiple harmonics. AM–PM distortion can dominate the out-of-band interference in a radio. At lower power levels, the phase deviation of the signal is approximately linear and the slope of the deviation, referred to as the modulation index, is often used as a figure-of-merit for the characterization of this nonlinearity. The *modulation index* is measured in degrees per volt using a VNA. The phase deviation is typically measured at the 1 dB compression point in order to determine modulation index. Because the VNA measures power, the computation of modulation index, k_ϕ, uses the formula

$$k_\phi = \frac{\Delta\Phi(P_{1\mathrm{dB}})}{2Z_0\sqrt{P_{1\mathrm{dB}}}} \tag{1.9}$$

where $\Delta\Phi(P_{1\mathrm{dB}})$ is the phase deviation from small signal at the 1 dB compression point, Z_0 is the characteristic impedance of the system and $P_{1\mathrm{dB}}$ is the 1 dB output compression point.

1.4.2.5 Adjacent Channel Power Ratio

Amplitude and phase distortion affect digitally modulated signals resulting in gain compression and phase deviation. The resulting signal, however, is far more complex than the simple one or two carrier results presented in Sections 1.4.2.2 through 1.4.2.4. Instead of IMD, *adjacent channel power ratio* (ACPR) is often specified for digitally modulated signals. ACPR is a measure of how much power leaks into adjacent channels of a radio due to the nonlinearities of the digitally modulated signal in a central channel. Measurement of ACPR is similar to measurement of IMD, but utilizes an appropriately modulated digital test signal in place of a single tone signal generator. Test signals for digitally modulated signals are synthesized using an *arbitrary waveform generator*. The output spectrum of the DUT in the channels adjacent to the tested channel are then monitored and power levels are measured.

1.4.2.6 Error Vector Magnitude

Adjacent channel power specifications are not adequate for certain types of modern digitally modulated systems. *Error vector magnitude* (EVM) is used in addition to, or instead of adjacent channel power for these systems. EVM specifications have already been written into system standards for GSM, NADC, and PHS, and they are poised to appear in many important emerging standards.

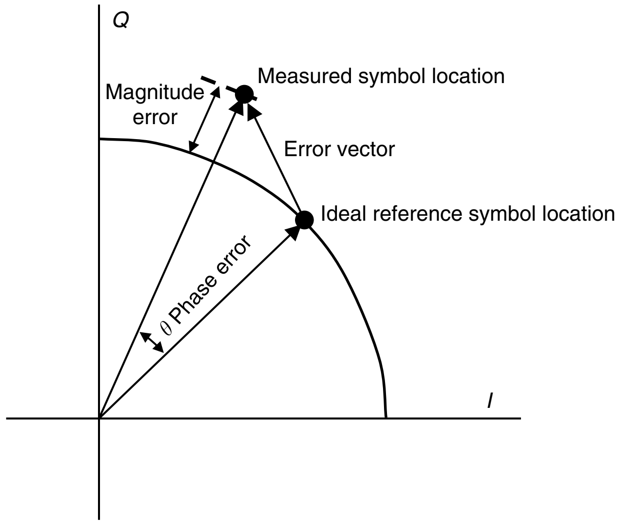


FIGURE 1.9 *I*–*Q* diagram indicating the error vector for EVM measurements.

The EVM measurement quantifies the performance of a radio transmitter against an ideal reference. A signal sent by an ideal transmitter would have all points in the *I*–*Q* constellation fall precisely at the ideal locations (i.e., magnitude and phase would be exact). Nonideal behavior of the transmitter, however, causes the actual constellation points to fall in a slightly scattered pattern that only approximates the ideal *I*–*Q* location. EVM is a way to quantify how far the actual points are from the ideal locations. This is indicated in Figure 1.9.

Measurement of EVM is accomplished using a vector signal analyzer (VSA). The equipment demodulates the received signal in a similar way to the actual radio demodulator. The actual *I*–*Q* constellation can then be measured and compared to the ideal constellation. EVM is calculated as the ratio of the root mean square power of the error vector to the RMS power of the reference.

1.4.3 Noise

Noise is a random process that can have many different sources such as thermally generated resistive noise, charge crossing a potential barrier, and generation–recombination (G–R) noise. Understanding noise is important in microwave systems because background noise levels limit the sensitivity, dynamic range and accuracy of a radio or radar receiver.

1.4.3.1 Noise Figure

At microwave frequencies noise characterization involves the measurement of noise power. The noise power of a linear device can be considered as concentrated at its input as shown in Figure 1.10. The figure considers an amplifier, but the analysis is easily generalized to other linear devices.

All of the amplifier noise generators can be lumped into an equivalent noise temperature with an equivalent input noise power per Hertz of $N_e = kT_e$, where k is Boltzmann's constant and T_e is the equivalent noise temperature. The noise power per Hertz available from the noise source is $N_S = kT_S$ as shown in Figure 1.10. Since noise limits the system sensitivity and dynamic range, it is useful to examine noise as it is related to signal strength using a signal-to-noise ratio (SNR). A figure-of-merit for an amplifier, *noise factor* (F), describes the reduction in SNR of a signal as it passes through the linear device illustrated in Figure 1.10. The noise factor for an amplifier is derived from the figure to be

$$F = \frac{\mathrm{SNR_{IN}}}{\mathrm{SNR_{OUT}}} = 1 + \frac{T_e}{T_S} \tag{1.10}$$

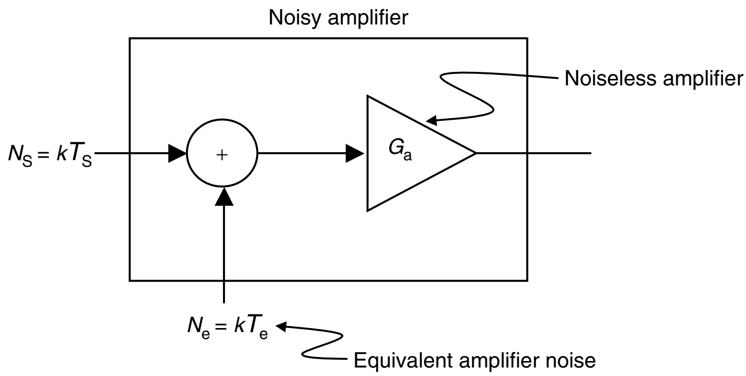


FIGURE 1.10 System view of amplifier noise.

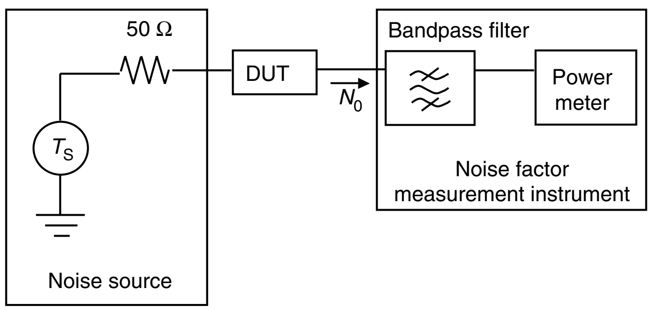


FIGURE 1.11 Measurement configuration for noise factor measurement.

Device noise factor can be measured as shown in Figure 1.11. To make the measurement, the source temperature is varied resulting in variation in the device noise output, N_0. The device noise contribution, however, remains constant. As T_S changes the noise power measured at the power meter changes providing a method to compute noise output.

In practice, the noise factor is usually given in decibels and called the *noise figure*,

$$\mathrm{NF} = 10 \log F \tag{1.11}$$

1.4.3.2 Phase Noise

When noise is referenced to a carrier frequency it modulates that carrier and causes amplitude and phase variations known as phase noise. Oscillator phase modulation (PM) noise is much larger than amplitude modulation (AM) noise. The phase variations caused by this noise result in *jitter* which is critical in the design and analysis of digital communication systems.

Phase noise is most easily measured using a spectrum analyzer. Figure 1.12 shows a typical oscillator source spectrum as measured directly on a spectrum analyzer. Characterization and analysis of phase noise is often described in terms of the power ratio of the noise at specific distances from the carrier frequency. This is illustrated in Figure 1.12.

1.4.4 Pulsed *I–V*

Although most of the measurements commonly utilized in RF and microwave engineering are frequency domain measurements, pulsed measurements are an important exception used to characterize

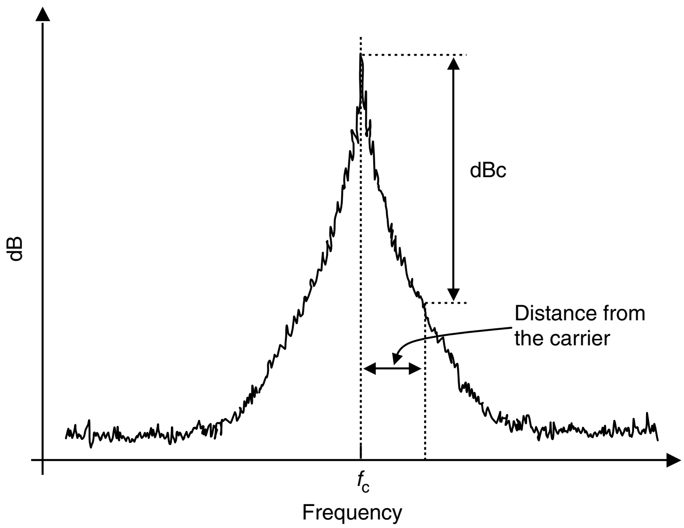


FIGURE 1.12 Typical phase noise spectrum observed on a spectrum analyzer.

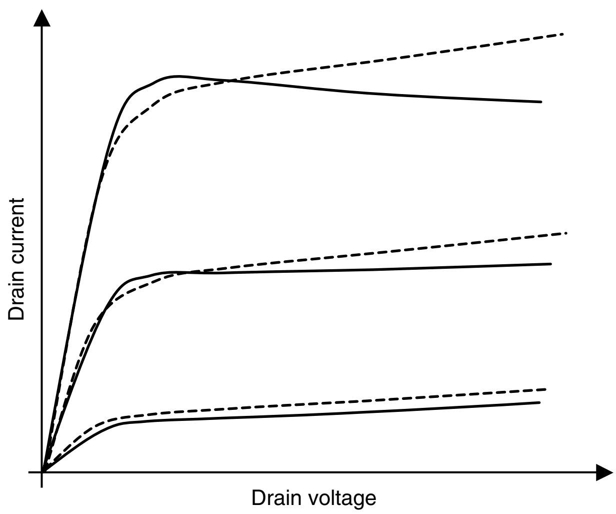


FIGURE 1.13 Pulsed *I–V* characteristics of a microwave FET. Solid lines are DC characteristics while dashed lines are pulsed.

high-frequency transistors. At RF and microwave frequencies, mechanisms known as *dispersion effects* become important to transistor operation. These effects reveal themselves as a difference in *I–V* characteristics obtained using a slow sweep as opposed to *I–V* characteristics obtained using a rapid pulse. The primary physical causes of *I–V* dispersion are thermal effects and carrier traps in the semiconductor. Figure 1.13 illustrates the characteristics of a microwave transistor under DC (solid lines) and pulsed (dashed lines) stimulation. In order to characterize dispersion effects, pulse rates must be shorter than the thermal and trapping time constants that are being monitored. Typically, for microwave transistors, that requires a pulse on the order of 100 ns or less. Similarly, the quiescent period between pulses should be long compared to the measured effects. Typical quiescent periods are on the order of 100 ms or more. The discrepancy between DC and pulsed characteristics is an indication of how severely the semiconductor traps and thermal effects will impact device performance.

Another use for pulsed I–V measurement is the characterization of high power transistors. Many high power transistors (greater than a few dozen Watts) are only operated in a pulsed mode or at a bias level far below their maximum currents. If these devices are biased at higher current levels for a few milliseconds, the thermal dissipation through the transistor will cause catastrophic failure. This is a problem for transistor model development, since a large range of I–V curves—including high current settings—is needed to extract an accurate model. Pulsed I–V data can provide input for model development while avoiding unnecessary stress on the part being characterized.

1.5 Circuits and Circuit Technologies

Figure 1.14 illustrates a generalized radio architecture that is typical of the systems used in many wireless applications today. The generalized diagram can apply to either communications or radar applications. In a wired application, the antenna of Figure 1.14 can be replaced with a transmission line. The duplexer of Figure 1.14 will route signals at the transmission frequency from the PA to the antenna while isolating that signal from the low noise amplifier (LNA). It will also route signals at the receive frequency from the antenna to the LNA. For some systems, input and output signals are separated in time instead of frequency. In these systems, an RF switch is used instead of a duplexer. Matching elements and other passive frequency selective circuit elements are used internally to all of the components shown in the figure. In addition, radio specifications typically require the use of filters at the ports of some of the components illustrated in Figure 1.14.

A signal received by the antenna is routed via the duplexer to the receive path of the radio. An LNA amplifies the signal before a mixer downconverts it to a lower frequency. The downconversion is accomplished by mixing the received signal with an internally generated local oscillator (LO) signal. The ideal receiver rejects all unwanted noise and signals. It adds no noise or interference and converts the signal to a lower frequency that can be efficiently processed without adding distortion.

On the transmitter side, a modulated signal is first upconverted and then amplified by the PA before being routed to the antenna. The ideal transmitter boosts the power and frequency of a modulated signal to that required for the radio to achieve communication with the desired receiver. Ideally, this

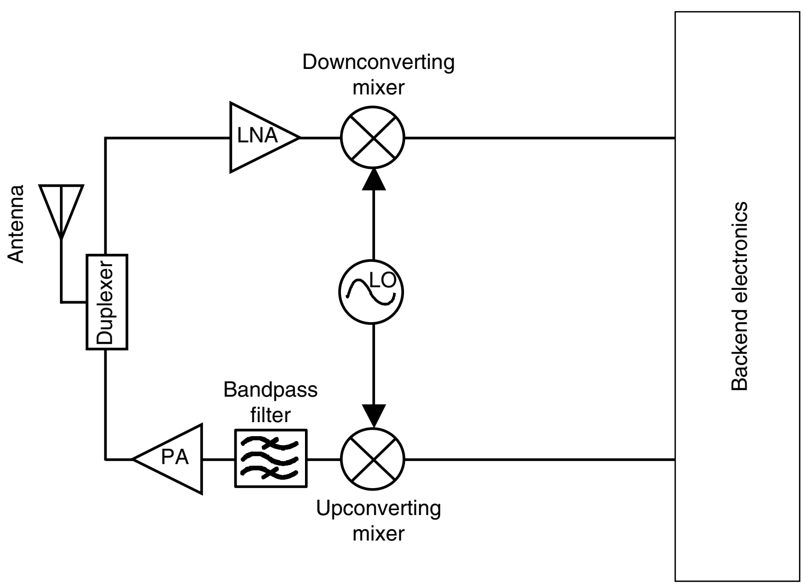


FIGURE 1.14 Generalized microwave radio architecture illustrating the microwave components in both the receiver and transmitter path.

process is accomplished efficiently (minimum DC power requirements) and without distortion. It is especially important that the signal broadcast from the antenna include no undesirable frequency components.

To accomplish the required transmitter and receiver functions, RF and microwave components must be developed either individually or as part of an integrated circuit. The remainder of this section will examine issues related to individual components that comprise the radio.

1.5.1 Low Noise Amplifier

The LNA is often most critical in determining the overall performance of the receiver chain of a wireless radio. The noise figure of the LNA has the greatest impact of any component on the overall receiver noise figure and receiver sensitivity. The LNA should minimize the system noise figure, provide sufficient gain, minimize nonlinearities, and assure stable 50 Ω impedance with low power consumption. The two performance specifications of primary importance to determine LNA quality are gain and noise figure.

In many radios, the LNA is part of a single chip design that includes a mixer and other receiver functions as well as the LNA. In these applications, the LNA may be realized using Silicon, SiGe, GaAs or another semiconductor technology. Si BJTs and SiGe HBTs dominate the LNA business at frequencies below a couple of GHz because of their tremendous cost and integration advantages over compound semiconductor devices. Compound semiconductors are favored as frequency increases and noise figure requirements decrease. For applications that require extremely low noise figures, cooled compound semiconductor HEMTs are the favored device.

1.5.2 Power Amplifier

A PA is required at the output of a transmitter to boost the signal to the power levels necessary for the radio to achieve a successful link with the desired receiver. PA components are almost always the most difficult and expensive part of microwave radio design. At high power levels, semiconductor nonlinearities such as breakdown voltage become critical design concerns. Thermal management issues related to dissipating heat from the RF transistor can dominate the design effort. Efficiency of the amplifier is critical, especially in the case of portable radio products. PA efficiency is essential to obtain long battery lifetime in a portable product. Critical primary design specifications for PAs include output power, gain, linearity, and efficiency.

For many applications, PA components tend to be discrete devices with minimal levels of on-chip semiconductor integration. The unique semiconductor and thermal requirements of PAs dictate the use of unique fabrication and manufacturing techniques in order to obtain required performance. The power and frequency requirements of the application typically dictate what device technology is required for the PA. At frequencies as low as 800 MHz and power levels of 1 Watt, compound semiconductor devices often compete with Silicon and SiGe for PA devices. As power and frequency increase from these levels, compound semiconductor HBTs and HEMTs dominate in this application. Vacuum tube technology is still required to achieve performance for some extremely high-power or high-frequency applications.

1.5.3 Mixer

A mixer is essentially a multiplier and can be realized with any nonlinear device. If at least two signals are present in a nonlinear device, their products will be produced at the device output. The mixer is a frequency translating device. Its purpose is to translate the incoming signal at frequency, f_{RF}, to a different outgoing frequency, f_{IF}. The LO port of the mixer is an input port and is used to *pump* the RF signal and create the IF signal.

Mixer characterization normally includes the following parameters:

- Input match (at the RF port)
- Output match (at the IF port)

- LO to RF leakage (from the LO to RF port)
- LO to IF leakage (from LO to IF port)
- Conversion Loss (from the RF port to the IF port)

The first four parameters are single frequency measurements similar to s-parameters S_{11}, S_{22}, S_{13}, and S_{23}. Conversion loss is similar to s-parameter S_{21}, but is made at the RF frequency at the input port and at the IF frequency at the output port.

Although a mixer can be made from any nonlinear device, many RF/microwave mixers utilize one or more diodes as the nonlinear element. FET mixers are also used for some applications. As in the case of amplifiers, the frequency of the application has a strong influence on whether Silicon or compound semiconductor technologies are used.

1.5.4 RF Switch

RF switches are control elements required in many wireless applications. They are used to control and direct signals under stimulus from externally applied voltages or currents. Phones and other wireless communication devices utilize switches for duplexing and switching between frequency bands and modes.

Switches are ideally a linear device so they can be characterized with standard s-parameters. Since they are typically bi-directional, $S_{21} = S_{12}$. Insertion loss (S_{21}) and reflection coefficients (S_{11} and S_{22}) are the primary characteristics of concern in an RF switch. Switches can be reflective (high impedance in the off state) or absorptive (matched in both on-and off-state).

The two major classes of technologies used to implement switches are PIN diodes and FETs. PIN diode switches are often capable of providing superior RF performance to FET switches but the performance can come at a cost of power efficiency. PIN diodes require a constant DC bias current in the on state while FET switches draw current only during the switching operation. Another important emerging technology for microwave switching is the micro-electro-mechanical systems (MEMS) switch. These integrated circuit devices use mechanical movement of integrated features to open and close signal paths.

1.5.5 Filter

Filters are frequency selective components that are central to the operation of a radio. The airwaves include signals from virtually every part of the electromagnetic spectrum. These signals are broadcast using various modulation strategies from TV and radio stations, cell phones, base stations, wireless LANs, police radar, and so on. An antenna at the front end of a radio receives all these signals. In addition, many of the RF components in the radio are nonlinear, creating additional unwanted signals within the radio. In order to function properly, the radio hardware must be capable of selecting the specific signal of interest while suppressing all other unwanted signals. Filters are a critical part of this selectivity. An ideal filter would pass desired signals without attenuation while suppressing signals at all other frequencies to elimination.

Although Figure 1.14 shows only one filter in the microwave portion of the radio, filters are typically required at multiple points along both the transmit and receive signal paths. Further selectivity is often accomplished by the input or output matching circuitry of amplifier or mixer components.

Filter characteristics of interest include the bandwidth or passband frequencies, the insertion loss within the passband of the filter, as well as the signal suppression outside of the desired band. The quality factor, Q, of a filter is a measure of how sharply the performance characteristics transition between passband and out-of-band behavior.

At lower frequencies, filters are realized using lumped inductors and capacitors. Typical lumped components perform poorly at higher frequencies due to parasitic losses and stray capacitances. Special manufacturing techniques must be used to fabricate lumped inductors and capacitors for microwave applications. At frequencies above about 5 or 6 GHz, even specially manufactured lumped element components are often incapable of producing adequate performance. Instead, a variety of technologies are exploited to accomplish frequency selectivity. Open- and short-circuited transmission line segments are often realized in stripline, microstrip, or coplanar waveguide forms to achieve filtering. Dielectric resonators, small pucks of high dielectric material, can be placed in proximity to transmission lines to achieve frequency selectivity. Surface acoustic wave (SAW) filters are realized by coupling the electromagnetic signal into piezoelectric materials and tapping the resulting surface waves with appropriately spaced contacts. Bulk acoustic wave (BAW) filters make use of acoustic waves flowing vertically through bulk piezoelectric material. MEMS are integrated circuit devices that combine both electrical and mechanical components to achieve both frequency selectivity and switching.

1.5.6 Oscillator

Oscillators deliver power either within a narrow bandwidth, or over a frequency range (i.e., they are tunable). Fixed oscillators are used for everything from narrowband power sources to precision clocks. Tunable oscillators are used as swept sources for testing, FM sources in communication systems, and the controlled oscillator in a phase-locked loop (PLL). Fixed tuned oscillators will have a power supply input and the oscillator output, while tunable sources will have one or more additional inputs to change the oscillator frequency. The output power level, frequency of output signal and power consumption are primary characteristics that define oscillator performance. The quality factor, Q, is an extremely important figure-of-merit for oscillator resonators. Frequency stability (jitter) and tunability can also be critical for many applications.

The performance characteristics of an oscillator depend on the active device and resonator technologies used to fabricate and manufacture the component. Resonator technology primarily affects the oscillator's cost, phase noise (jitter), vibration sensitivity, temperature sensitivity, and tuning speed. Device technology mainly affects the oscillator maximum operating frequency, output power, and phase noise (jitter).

Resonator choice is a compromise of stability, cost, and size. Generally the quality factor, Q, is proportional to volume, so cost and size tend to increase with Q. Technologies such as quartz, SAW, yttrium-iron-garnet (YIG) and dielectric resonators allow great reductions in size while achieving high Q by using acoustic, magnetic, and dielectric materials, respectively. Most materials change size with temperature, so temperature stable cavities have to be made of special materials. Quartz resonators are an extremely mature technology with excellent Q, temperature stability, and low cost. Most precision microwave sources use a quartz crystal to control a high-frequency tunable oscillator via a PLL. Oscillator noise power and jitter are inversely proportional to Q^2, making high resonator Q the most direct way to achieve a low noise oscillator.

Silicon bipolar transistors are used in most low noise oscillators below about 5 GHz. Heterojunction bipolar transistors (HBTs) are common today and extend the bipolar range to as high as 100 GHz. These devices exhibit high gain and superior phase noise characteristics over most other semiconductor devices. For oscillator applications, CMOS transistors are poor performers relative to bipolar transistors, but offer levels of integration that are superior to any other device technology. Above several GHz, compound semiconductor MESFETs and HEMTs become attractive for integrated circuit applications. Unfortunately, these devices tend to exhibit high phase noise characteristics when used to fabricate oscillators. Transit time diodes are used at the highest frequencies where a solid-state device can be used. IMPATT and Gunn diodes are the most common types of transit-time diodes available. The IMPATT diode produces power at frequencies approaching 400 GHz, but the avalanche breakdown mechanism inherent to its operation causes the device to be very noisy. In contrast, Gunn diodes tend to exhibit very clean signals at frequencies as high as 100 GHz.

1.6 CAD, Simulation, and Modeling

The unique requirements of RF and microwave engineering establish a need for design and analysis tools that are distinct from conventional electrical engineering tools. Simulation tools that work well for a digital circuit or computational system designer fail to describe RF and microwave behavior adequately. Component and device models must include detailed descriptions of subtle parastic effects not required for digital and low-frequency design. Circuit CAD tools must include a much wider range of components than for traditional electrical circuit design. Transmission line segments, wires, wire bonds, connector transitions, specialized ferrite, and acoustic wave components are all unique to microwave circuit design. In addition, the impact of the particular package technology utilized as well as layout effects must be considered by the microwave engineer. Electromagnetic simulators are also often required to develop models for component transitions, package parasitics, and complex board layouts.

In the microwave design environment, a passive component model for a single chip resistor requires a model that uses several ideal circuit elements. Shunt capacitances are required to model parasitic displacement currents at the input and output of the component. Ideal capacitors, inductors, or transmission line segments must be used internal to the desired component to model phase-shift effects. Nonideal loss mechanisms require the inclusion of additional resistor elements. Similar complexities are required for chip capacitors and inductors. The complexity is compounded by the fact that the ideal element values required to model a single component will change depending on how that component is connected to the circuit board and the kind of circuit board used.

Device models required for microwave and RF design are significantly more detailed than those required for digital or low-frequency applications. In digital design, designers are concerned primarily with two voltage–current states and the overall transition time between those states. A device model that approximately predicts the final states and timing is adequate for many of these applications. For analog applications, however, a device model must describe not only the precise I–V behavior of the device over all possible transitions, but also accurately describe second and third derivatives of those transitions for the model to be able to predict second-and third-order harmonics. Similar accuracy requirements also apply to the capacitance-voltage characteristics of the device. In the case of PA design, the device model must also accurately describe gate leakage and breakdown voltage—effects that are considered second order for many other types of circuit design. Development of a device model for a single microwave transistor can require weeks or months of detailed measurement and analysis.

Another characteristic that distinguishes RF microwave design is the significant use of electromagnetic simulation. A microwave circuit simulation that is completely accurate for a set of chips or components in one package may fail completely if the same circuitry is placed in another package. The transmission line-to-package transitions and proximity effects of package walls and lid make these effects difficult to determine and model. Often package effects can only be modeled adequately through the use of multi-dimensional electromagnetic simulations. Multilevel circuit board design also requires use of such simulations. Radiation effects can be important contributors of observed circuit behavior but cannot be captured without the use of electromagnetic simulation.

Because of the detail and complexity required to perform adequate RF and microwave design, the procedure used to develop such circuits and systems is usually iterative. Simple ideal-element models and crude models are first used to determine preferred topologies and approximate the final design. Nonideal parasitics are then included and the design is reoptimized. Electromagnetic simulators are exercised to determine characteristics of important transitions and package effects. These effects are then modeled and included in the simulation. For some circuits, thermal management can become a dominant concern. Modeling this behavior requires more characterization and simulation complexity. Even after all of this characterization and modeling has taken place, most RF and microwave circuit design efforts require multiple passes to achieve success.

References

1. Collin, R. E., *Foundations for Microwave Engineering*, McGraw-Hill, New York, 1992, p. 2.
2. Adam, S. F., *Microwave Theory and Applications*, Prentice Hall, Englewood Cliffs, NJ, 1969.
3. Halliday, D., Resnick, R. and Walker, J., *Fundamentals of Physics*, 7th ed., John Wiley & Sons, Hoboken, NJ, 2005.

4. Schroder, D. K., *Semiconductor Material and Device Characterization*, John Wiley & Sons, Hoboken, NJ, 1990.
5. Plonus, M. A., *Applied Electromagnetics*, McGraw-Hill, New York 1978.
6. Jonscher, A. K., *Dielectric Relaxation in Solids*, Chelsea Dielectric Press, London, 1983.
7. Kittel, C., *Introduction to Solid State Physics*, 3rd ed. John Wiley & Sons, Hoboken, NJ, 1967.
8. Tsui, J. B., *Microwave Receivers and Related Components*, Peninsula Publishing, Los Altos Hills, CA, 1985.
9. Pearce, C. W. and Sze, S. M., *Physics of Semiconductors*, John-Wiley & Sons, New York, 1981.
10. Bahl, I. and Bhartia P., *Microwave Solid State Circuit Design*, John Wiley & Sons, New York, 1988.
11. Lee, W. C. Y., *Mobile Communications Engineering*, McGraw-Hill Book Company, New York, 1982.
12. Jakes, W. C., Jr. (ed.), *Microwave Mobile Communications*, John Wiley & Sons, Inc., New York, 1974.
13. Evans, J., *Network Interoperability Meets Multimedia*, Satellite Communications, pp. 30–36, February 2000.
14. Laverghetta, T. S., *Modern Microwave Measurements and Techniques*, Artech House, Dedham, MA, 1989.
15. Ramo, S., Winnery, J. R., and Van Duzer, T., *Fields and Waves in Communications Electronics*, 2nd ed., John Wiley & Sons, New York, 1988.
16. Matthaei, G. L., Young, L., and Jones, E. M. T., *Microwave Filters, Impedance-Matching Networks, and Coupling Structures*, Artech House, Dedham, MA, 1980.
17. Ambrozy, A., *Electronic Noise*, McGraw-Hill, New York, 1982.
18. Watson, H. A., *Microwave Semiconductor Devices and their Circuit Applications*, McGraw-Hill, New York, 1969.
19. Vendelin, G., Pavio, A. M., and Rohde, U. L., *Microwave Circuit Design*, John Wiley & Sons, Hoboken, NJ, 1990.

I

Microwave and RF Product Applications

2

Terrestrial and Satellite Mobile Radio Systems

Andy D. Kucar
www.radio4u.com

2.1 Introduction

Continuous developments and technological advances in Silicon Germanium (SiGe), Silicon Carbide (SiC), Gallium Arsenide (GaAs), Indium Phosphide (InP), insulation and substrate materials, microwave monolithic integrated circuit (MMIC), application specific integrated circuit (ASIC), System on Chip (SoC), System in Package (SiP), microprocessors, analog/digital signal processing (A/DSP), and battery technology, supported by computer-aided design (CAD) and robotics manufacturing allow a viable implementation of miniature radio transceivers at radio frequencies as high as 100 GHz, that is, at wavelengths as short as about 3 mm. At frequency bands up to about 6 GHz, additional spectra have been assigned to mobile services, see Figure 2.1; corresponding shorter wavelengths allow a viable implementation of adaptive antennas and diversity schemes empowered with powerful spatial turbo coding schemes; these Multiple Input multiple output (MIMO) systems are capable of offering higher quality of transmission and higher spatial frequency spectrum efficiency.

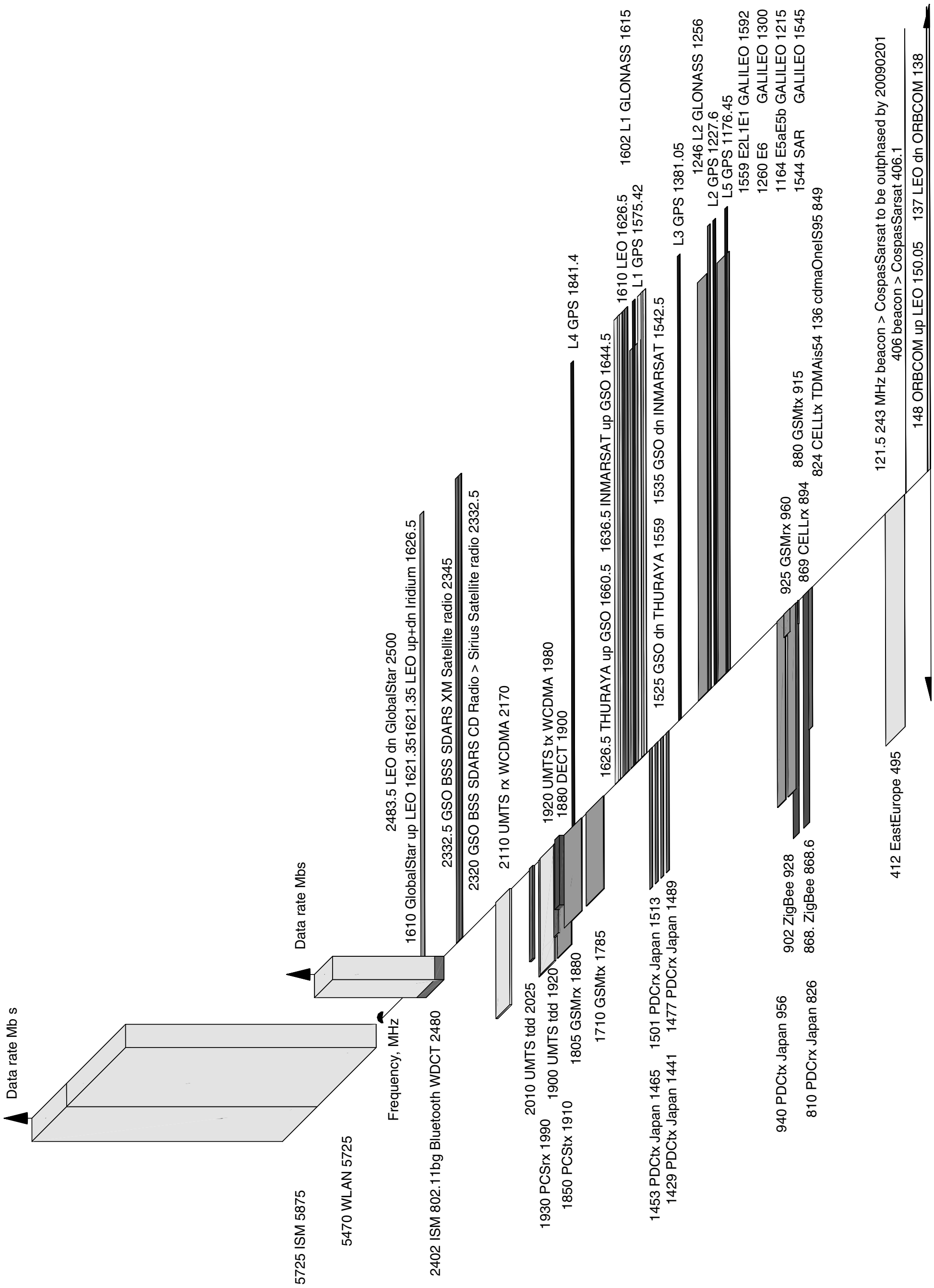


FIGURE 2.1 Mobile systems: data rate versus frequency and distance.

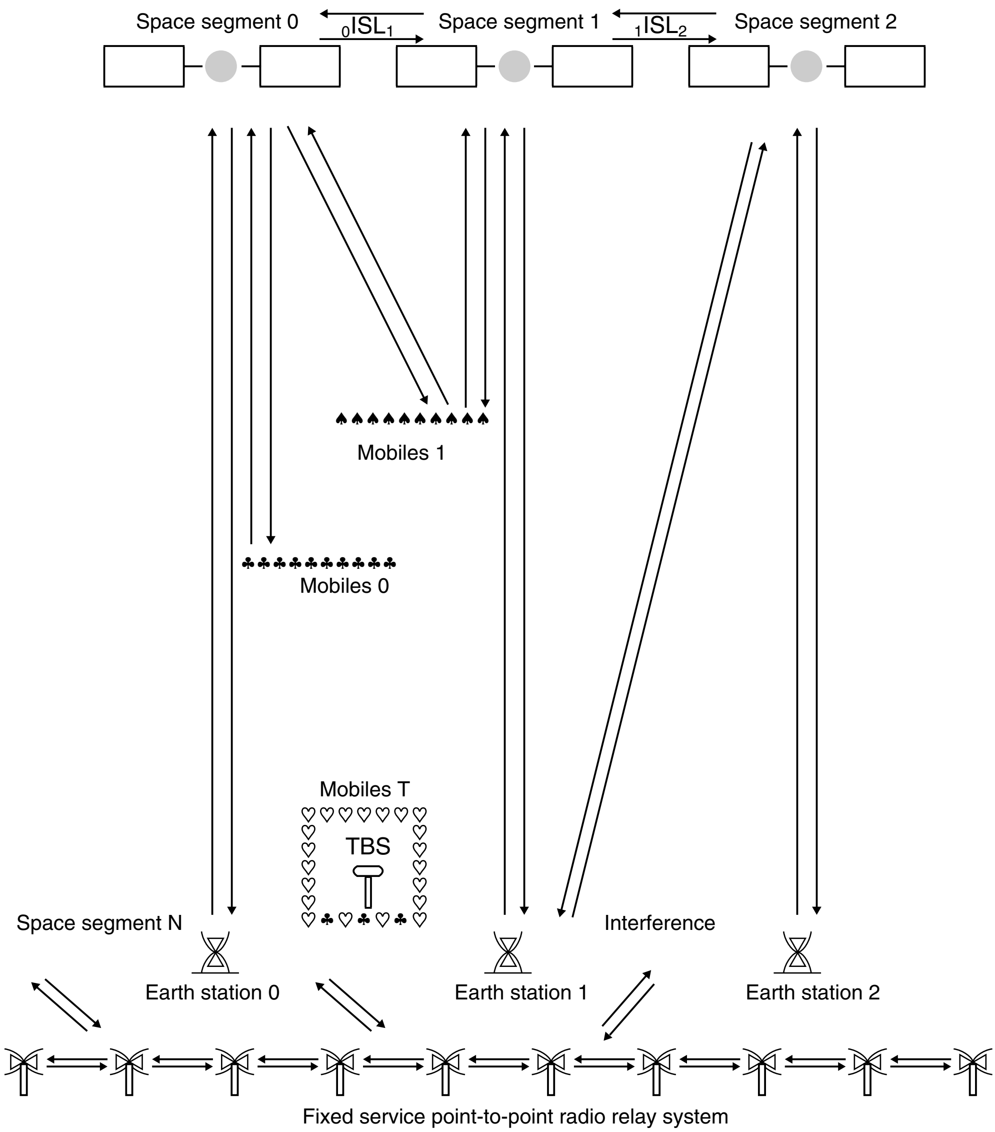


FIGURE 2.2 A model of fixed and mobile, satellite, and terrestrial radio systems.

The turbulent flux of market forces (excited by the possibilities of a myriad of new services and great profits), international and domestic standard forces (that manage common natural resource—the airwaves), and technology forces (capable of creating viable products) for some time acted harmoniously and created a broad choice of communications (voice and data), information, and navigation systems, which propelled an explosive growth of mobile radio services for travellers; recently, profiteers have decimated western telecommunications industries.

The terrestrial and satellite mobile radio communications, radionavigation and some audio broadcasting systems addressed in this contribution, and illustrated in Figure 2.2, include:

- The first generation analog cellular mobile radio systems such as North American *Advanced Mobile Phone Service* (AMPS), *Japanese Mobile Communications System* (MCS), Scandinavian *Nordic Mobile Telephony system* (NMT) and British *Total Access Communication System* (TACS) used analog voice data and frequency modulation (FM) for the transmission of voice, and coded digital data and frequency shift keying (FSK) modulation scheme for the transmission of control information. Conceived and designed in the 1970s, these systems were optimized for professional vehicle based services such as police and ambulances operating at possibly high vehicle speeds and the first generation analog *Cordless Telephones* (CT) including CT0 and CT1 cordless telephone systems that are intended for use in the household environment. Worldwide, most, but not all, of the first generation systems have been upgraded to the second and higher generation systems.

- The second generation digital cellular and personal mobile radio systems include *Global System for Mobile Communications* (GSM), *Digital* AMPS > IS-54/136, *Digital Cellular System* (DCS) 1800/1900 and *Personal Digital Cellular* (PDC), all Time Division Multiple Access (TDMA) and IS-95 *spread spectrum Code* Division Multiple Access (CDMA) system, recently branded as cdmaOne. All mentioned systems employ digital data for both voice and control purposes. Other second generation systems are *digital Cordless Telephony systems* CT2, CT2Plus and CT3, *Digital Enhanced Cordless Telephone* (DECT) and *Personal Handyphone System* (PHS); see Table 2.1. Wireless data mobile radio systems include ARDIS, RAM, *Trans European* (*TErrestrial*) *Trunked RAdio* (TETRA), which seem to be merging into the *General* (*Global*) *Packet Radio Service* (GPRS) and even into the GSM system, *Cellular Digital Packet Data* (CDPD), IEEE 802.11 *Wireless Local Area Network* (WLAN) and Bluetooth. There are also projects known as *Personal Communications Network* (PCN), *Personal Communications Systems* (PCS) and, under the umbrella of ITU, the *Future Public Land Mobile Telecommunications Systems* (FPLMTS) > *Universal Mobile Telecommunications Systems* (UMTS) > *International Mobile Telecommunications system* (IMT) −2000 > 3G > 4G project. Here 3G and 4G stand for the third and the fourth generation systems, respectively, 3GPP, WiFi, WiMAX, ZigBee, and so on.

A number of manufacturers and institutions have been supporting diverse interest groups, domestic and international organizations, with the aim of selling their own ideas and products, in fact creating the industry standards; requirements, not necessarily unique and consistent, and by no means cast in stone, have been set for future generations of mobile wireless systems, and those of 3G and 4G in particular. Generally, a system of a higher generation should be able to provide a wider bandwidth, which would facilitate offering of wider scope of services, including video broadcasting and exchange. A wider bandwidth essentially means wider profit potentials; yet, technically and economically, a wide bandwidth, that is, a broadband system is more difficult to realize on a global scale. In the meantime, fearsome and costly battles, causing enormous collateral damage, have been fought for customers pockets and loyalty with old and new, often half backed, ideas, slogans, acronyms (see Table 2.2), and products; some of those products and/or systems are loosely classified into

- 2.5G includes GPRS and *Wideband Integrated Dispatch Enhanced Network* (WiDEN).
- 2.75G includes CDMA2000 *Radio Transmission Technology* CDMA2000 1 × RTT and *Enhanced Data rates for Global* (GSM) *Evolution* EDGE.
- 3G includes *Wideband* CDMA W-CDMA, UMTS, *Freedom of Mobile Multimedia Access* (FOMA), CDMA2000 *Evolution* (CDMA2000 1 × EV), and *Time Division Synchronous Code Division Multiple Access* (TD-SCDMA).
- 3.5G includes *High Speed Downlink Packet Access* (HSDPA).
- 3.75G includes *High Speed Uplink Packet Access* (HSUPA).
- The first generation of a commercial satellite system introduced by INMARSAT, provided analog voice, facsimile and low speed data to commercial navy ships equipped with large gimbal mounted antennas, worldwide. The second generation INMARSAT systems concentrated on a digitalization and expansion of low data rate services to smaller ships and yachts. In addition, other entities started to include new satellite services via payloads piggybacked on fixed service (FS) satellites, or on dedicated satellites for mobile users. These include *satellite mobile radio systems* such as OmniTRACS, North American *Mobile SATellite* system (MSAT), *AUStralian SATellite* system (AUSSAT) > Optus, Iridium, Globalstar, the most advanced more than 200 spot beams Thuraya (with characteristics of a 3G or even a 4G system), two radio broadcasting satellite systems Sirius and XM

TABLE 2.1 The Comparison of Cordless Telephone (CT), CDMA, and PHT Systems

	System name							
Parameter	CT0	CT1/+	JCT	CT2/+	CT3	DECT	CDMA	PHT
TX frequency (MHz)								
Base	22,26,30,31,46,48,45	914/885	254	864–868, 944–948	944–948	1880–1900		1895–1907
Mobile	48,41,39,40,69,74,48	960/932	380	864–868, 944–948	944–948	1880–1900		1895–1907
Multiple access method	FDMA	FDMA	FDMA	F/TDMA	TDMA	TDMA	CDMA	F/TDMA
Duplexing method	FDD	FDD	FDD	TDD	TDD	TDD	FDD	TDD
Channel spacing (kHz)	1.7,20,25,40	25	12.5	100	1000	1728	1250	300
Channel rate (kb/s)				72	640	1152	1228.80	384
Channels/RF	1	1	1	1	8	12	32	4
Channels/band	10,12,15,20,25	40,80	89	20	2	5		20
—BotStrut Burst/frame length (ms)				1/2	1/16	1/10	n/a	1/5
Modulation type	FM	FM	FM	GFSK	GMSK	GMSK	B/QPSK	$\pi/4$
Coding				Cyclic, RS	CRC 16	CRC 16	Conv 1/2, 1/3	
Transmit power (mW)				$\leq$10	$\leq$80	$\leq$100	$\leq$10	$\leq$80
Transmit power steps				2	1	1	Many	Many
TX power range (dB)				16	0	0	$\geq$ 80	
Vocoder type	Analog	Analog	Analog	ADPCM	ADPCM	ADPCM	CELP	ADPCM
Vocoder rate (kb/s)				Fixed 32	Fixed 32	Fixed 32	$\leq$9.6	Fixed 32
Max data rate (kb/s)				32	ISDN 144	ISDN 144	9.6	384
Processing delay (ms)	0	0	0	2	16	16	80	
[1] Minimum				1/25	1/15	1/15	1/4	
Average				1/15	1/07	1/07	2/3	
Maximum				[2] 1/02	[2] 1/02	[2] 1/02	3/4	
[3]				100 × 1	10 × 8	6 × 12	4 × 32	
[4] Minimum				4	5–6	5–6	[5] 32 (08)	
Average				7	11–12	11–12	85 (21)	
Maximum				[2] 50	[2] 40	[2] 40	96 (24)	

[1] Reuse efficiency.
[2] The capacity (in the number of voice channels) for a single isolated cell.
[3] Theoretical number of voice channels per cell and 10 MHz.
[4] Practical number of voice channels per 10 MHz.
[5] The capacity in parentheses may correspond to a 32 kbit/s vocoder.
Reuse efficiency and associate capacities reflect our own estimates.

Source: www.radio4u.com, 20061020.

cordless-sys

TABLE 2.2 Glossary of Terms

3GPP	3rd Generation Partnership Project
nG	n-th Generation cellular radio/wireless system
ACI	Adjacent Channel Interference
AGC	Automatic Gain Control
ALC	Automatic Level Control
ALOHA	Data protocol, also slotted ALOHA, greetings in Hawaii
AMPS	Advanced Mobile Phone Service
APC	Automatic Power Control (on transmitter side)
ARPANET	Advanced Research Project Agency Network, Internet's forefather
ARQ	Automatic Repeat reQuest
ASIC	Application Specific Integrated Circuits
BER	Bit Error Rate
C/I	Carrier-to-Interference ratio
C/N, CNR	Carrier-to-Noise Ratio
CAD	Computer Aided Design
CB	Citizen Band (mobile radio)
CCI	CoChannel Interference
CDMA	spread spectrum Code Division Multiple Access
CEPT	Conference of European Postal and Telecommunications (Administrations)
CIR, C/I	Carrier-to-Interference Ratio
COFDM	Coded Orthogonal Frequency Division Multiplexing modulation
COST	COoperation in the field of Scientific and Technical research, Europe
CTn	Cordless Telephony of generation $n = 0, 1, 2$
DAMPS	Digital AMPS, also Dual mode AMPS, IS-54 > IS-136
DOC	Department of Communications (in Canada)
DSP	Digital Signal Processing
EGNOS	European Geostationary Navigation Overlay Service
ETSI	European Telecommunications Standards Institute
FCC	Federal Communications Commission (in United States)
FDD	Frequency Division Duplex
FDMA	Frequency Division Multiple Access
FHMA	Frequency Hopping Multiple Access
FPLMTS	Future Public Land Mobile Telecommunications Systems
GALILEO	European radionavigation system proposal
GDSS	Global Distress Safety System
GLONASS	Global Navigation Satellite System, USSR > Russia
GNSS	Global Navigation Satellite System (ICAO)
GOES	Geostationary Operational Environmental Satellites
GPRS	General (Global) Packet Radio Service, data speeds $\leq$115 kbit/s
GPS	Global Positioning System
GSM	Groupe Spécial Mobile > now Global System for Mobile communications
HSDPA	High Speed Downlink Packet Access
HSUPA	High Speed Uplink Packet Access
IrDA	Infrared Data Association
IEEE	Institute of Electronic and Electrical Engineers, standard body
IMT	IMT-2000, FPLMTS > International Mobile Telecommunications, 3G
ISDN	Integrated Service Digital Network
ISM	Industrial Scientific Medical frequency band(s)
ITU	International Telecommunications Union
LAN	Local Area Network, see MAN, WAN
LORAN	LOng-RAnge Navigation system, www.gps.gov/loran/default.htm
MIMO	Multiple Input Multiple Output system
MMIC	Microwave Monolitic Integrated Circuits
MOS	Mean Opinion Score
NMC	Network Management Center
NMT	Nordic Mobile Telephone (system)
OFDMA	Orthogonal Frequency Division Multiplexing Access

TABLE 2.2 Continued

PCN	Personal Communications Networks
PCS	Personal Communications Systems
PSTN	Public Switched Telephone Network
QoS	Quality of Service
SiP	System in Package
SoC	System on Chip
SARSAT	Search And Rescue Satellite Aided Tracking system
SDMA	Space (Spatial) Division Multiple Access
SERES	SEarch and REscue Satellite
SS	Spread Spectrum
TACS	Total Access Communication System
TDMA	Time Division Multiple Access
UMTS	FPLMTS > Universal Mobile Telecommunications Systems
WiBro	Wireless Broadband, Korean
WiDEN	Wideband Integrated Dispatch Enhanced Network
WiFi	Wireless Fidelity IEEE 802.11
WiMAX	Worldwide Interoperability for Microwave Access, IEEE 802.16
WARC	World Administrative Radio Conference

Source: www.radio4u.com, 20061020.

Satellite Radio, and low rate data collection and exchange solution by Orbcomm. In the meantime, INMARSAT upgraded its fleet of satellites to include radionavigation services, privatized and introduced a huge multibeam generation of satellites with increased capacity, bandwidth, and data rate of up to 128 kbit/s.

After a brief prologue and historical overview, technical issues such as the repertoire of systems and services, the airwaves management, the operating environment, service quality, network issues and cell size, channel coding and modulation, speech coding, diversity, multiple broadcasting (FDMB, TDMB, CDMB), and multiple access (FDMA, TDMA, CDMA) are briefly discussed.

Many existing mobile radio communications systems collect some form of information on network behavior, users' positions, and so on, with the purpose of enhancing the performance of communications, improving handover procedures, and increasing the system capacity. Coarse positioning is usually achieved inherently, while more precise positioning and *navigation* can be achieved by employing LORAN-C and/or GPS, differential GPS, GLONASS, *Wide Area Augmentation System* (WAAS), *Multi-functional Transport SATellite* (MTSAT), *European Geostationary Navigation Overlay Service* (EGNOS), GALILEO, *Global Navigation Satellite System* (GNSS) signals, or some other means, at an additional, usually modest, increase in cost and complexity (Kucar, 2006b).

2.2 Prologue

Mobile radio systems provide their users with opportunities to travel freely within the service area and at the same time being able to communicate with any telephone, fax, data modem, Internet connection, and electronic mail subscriber anywhere in the world; to determine their own positions, to track the precious cargo, to improve the management of fleets of vehicles and the distribution of goods, to improve traffic safety, to provide vital communication links during emergencies, search and rescue operations, to browse their favorite web sites, and so on. These *tieless* (*wireless, cordless*) communications, the exchange of information, determination of position, course, and distance travelled are made possible by the unique property of the radio to employ an *aerial* (*antenna*) for radiating and receiving electromagnetic waves. When the user's radio antenna is stationary over a prolonged period of time, the term *fixed radio* is used; a radio transceiver capable of being carried or moved around, but stationary when in operation, is called *portable radio*; a radio transceiver capable of being carried and used, by a vehicle or by a person on the move, is called *mobile radio, personal and/or handheld device.* Individual radio users may communicate directly, or via one or more intermediaries, which may be *passive radio repeater*(*s*), *base station*(*s*), *or switch*(*es*). When all intermediaries are located on the Earth, the terms *terrestrial radio system* and *radio system* have been used; when at least one intermediary is satellite-borne, the terms *satellite radio system and satellite system* have been used. According to the location of a user, the terms *land*, *maritime*, *aeronautical*, *space*, and *deep space radio systems* have been used. The second unique property of all, terrestrial and satellite, radio systems is that they all share the same natural resource—the *airwaves* (*frequency bands and the space*).

2.3 A Glimpse of Telecommunications History

On 17910302 in Paris, (1763, Abbe Claude Chappe, 1805) demonstrated his *tachigraphe* (a quick writer), that is, a mechanical optical telegraph. An improved version of Chappe semaphore was used to create 23 stations and 230-km long Paris–Lille optical telegraph line. On 17940815 the first telegram was sent over the line to Paris reporting (1769, Napoleon(e) Bonapart(e) 1799 First Consul 1804 Emperor Napoleon I Le Grande 1814, 1815, 1821)'s victory in the field. Soon after, numerous optical telegraph lines sprung across Europe.

In 1798, after two years of experimentation, (1751, Francisco Salvá i Campillo, 1828) a doctor from Barcelona, and (1758, Augustin de Betancourt y Molyna, 1824) an engineer, a native of Santa Cruz de Tenerife, installed an electric telegraph between Madrid and Aranjuez, Spanish Empire; a volta pile metal battery and one wire per letter solution, as described in *Scots Magazine* by one C.M. on 17530217, were used.

On 18090828 at Bavarian Academy in Munich, (1755, Samuel Thomas von Sömmerring, 1830), Professor of Anatomy, demonstrated a five needle electrochemical telegraph to (1786, Paul Constadt Baron Pawel Lwowić Schilling, 1837), a Russian engineer. In 1832 near Petro(v)grad (St. Petersburg), a needle telegraph between Russian Emperor (1796, 1825 Nicholas I 1855, 1855)'s winter and summer palace was installed. Schilling improved von Sömmerring's telegraph, devised his own electromechanical version, and just before his premature death, acquired rights to install telegraph lines across Russia.

(1777, Johann Carl Friedrich Gauss, 1855), child prodigy, student and professor at Göttingen University, graduate of Helmstedt University, 1807 director of the Göttingen astronomical observatory, German scientist, mathematician, astronomer, geodetic and electric engineer, inventor, known as the Prince of Mathematicians, second only to Euler, the great man of science and of impeccable character, elevated many of previous topics to famous Gauss rigorous levels, established new laws and laid the foundations for a number of mathematical branches and sciences. In 1831, (1804, Wilhelm Eduard Weber, 1891), student and 1826 professor at University of Halle, German physicist, was appointed to the chair of physics at Göttingen University. Six years of close friendship and productive collaboration between Gauss and W. Weber followed. Together, they established a system of absolute physical units, still in use, which served as a basis for any new development. They also discovered *current and voltage circuit laws*, (re)discovered in 1847 and popularized by (1824, Gustav Robert Kirchhoff, 1887), German scientist and mathematical physicist. On 18321021, Gauss and W. Weber installed a 2.1-km long electromagnetic telegraph line at the University of Göttingen. The line was extended and inaugurated on 18330407 and was in operation until 18451216, before it was destroyed by lightning.

In 1836, (1801, Karl August Steinhall, 1870), Professor of Mathematics and Physics in Munich, Bavaria, constructed an electromagnetic telegraph/teleprinter. In 1838, he constructed a line between the Bavarian Academy in Munich and the astronomical observatory in Bogenhausen. On 18390709, (1806, William Fothergill Cooke, 1879), who as a student in Heidelberg was introduced to Schilling's telegraph, and his partner (1802, Charles Wheatstone, 1875), Heaviside's uncle, professor at Kings College in London, installed Liverpool–Manchester telegraph line, the first in England.

In 1832, on his return trip home from Paris to the United States aboard the sailing ship *Sully*, a painter (1791, Samuel Finley Breese Morse, 1872) heard about European telegraph experiments from a Bostonian engineer. Later on, Morse's subordinate (1807, Alfred Vail, 1859), engineer from New Jersey, studied existing telegraphs and came with his own much improved electromechanical telegraph/teleprinter and his own

MorseVailCode. On 18440524, Morse and Vail sent the first message over the Baltimore–Washington, D.C. telegraph line.

In 1842 in France, (1804, Louis François Clément Bréguet, 1883), the owner of famous family watch making business, the grandson of codesigner of *Chappe's telegraph* and famous clock maker (1747, Louis Abraham Bréguet, 1832), started manufacturing a telegraph equipment of his own design. In 1880, Louis François Clément Bréguet and his son (1851, Antoine Bréguet, 1882) expanded their production portfolio to the *Bell–Bréguet* telephone equipment. (1880, Louis Charles Bréguet, 1955), son of Antoine Bréguet, French aviation pioneer, engineer, and industrialist, developed a wind tunnel in 1905, in late summer of 1907, together with his brother (1881, Jacques Eugène Henri Bréguet, 1939) and Professor (1850, Charles Robert Richet, 1935) the *Gyroplane 1*; Bréguet first aircraft, a rugged biplane was designed in 1909; the company manufactured over 8000 *Bréguet 14* reconnaissance aircraft for the Allied Forces; in 1919, Louis Charles Bréguet established a commercial air transportation company, *Compagnie Des Messageries Avienne*, to evolve in *Air France*.

In 1848 in what is now Germany, (1816, Ernst Werner von Siemens, 1892), an engineer and businessman, and his company *Siemens & Halske* (*S&H*, established in 1847), installed a telegraph line between Berlin and Frankfurt on Main. On 18500928, (1801, 1848 Ban Josip Jelačić Bužinski, 1859) sent a telegram from Vienna to Agram (Zagreb), signaling the end of the Hungarian Revolution. In 1853–1855, *S&H* installed the telegraph lines in Russia, between Abo (now Finland) and Sevastopol (now Ukraine). In 1882, at the Paris Conference, an absolute unit for electrical conductivity, Siemens, was introduced. In 1869, (1845, Emile Baudot, 1903), an engineer, joined the *French Telegraph Administration*. He promptly improved then existing telegraph and teleprinter equipment. In 1877, his five unit Baudot Code, and a unit 1 Bd $\equiv$ 1 symbol/sec were introduced.

In 1849 in Havana, Cuba, then a part of the Spanish Empire, (1808, Antonio Santi Giuseppe Meucci, 1889), Florence, now part of Italy, born theater stage master technician, able chemist and physicist, entrepreneur, and inventor, designed a teletrophone, that is, a telephone. In 1850 he moved to New York City, where he used the teletrophones in his house and candle factory. On 18711228 Meucci applied for a caveat for his invention; later, he was severely burned and crippled in an explosion and hospitalized for a longer period. To cover hospital and leaving expenses, his wife sold Meucci's numerous teletrophone models for \$8. These teletrophones soon appeared in laboratories of *Bell*, *Western Electric*, and *Edison* Companies. In the 1870s and 1880s Meucci was sued and vilified in newspapers, while his patent records vanished; his temporary patent number 3,335 expired in 1874 due to, allegedly, lack of funds. Meucci's experiment performed on 18700927, and corresponding drawings, include inductive loading coils, patented by (1858, Mihajlo Michael Idvorski Pupin, 1935) on 19000619. In 2002, the U.S. Administration declared Meucci to be the first inventor of the telephone; (in)justice mills do grind slowly.

On 18611026 in Frankfurt, Germany, (1834, Johann Philip(p) Reis, 1874), a school master from Friedrichsdorf Germany, demonstrated an apparatus for electrical reproduction and transmission of sound, called telephone. He produced and sold about dozen of telephone devices but was not able to commercialize his product any further. On 18760307 (1847, Alexander Graham Bell, 1922), Scotland born Canadian then American teacher and engineer, received the U.S. patent 174,465 on the subject of *Improvement in Telegraphy*, telephone was neither mentioned nor described in the application. Bell's patent application preceded the one of (1835, Elisha Gray, 1901), cofounder of *Western Electric*, by minutes; later, at least one employee of the Patent Office was jailed for fraud. On 18770427 Edison filed a patent application for an improved telephone transmitter. On 18770709 the *Bell Telephone Company* (*BTC*) was founded; on 18991230 BTC became the *American Telephone & Telegraph Company AT&T*. In 1870s, it was reported that the Thumtsein (telephone) was already invented in the year 968 AD by the Chinese inventor Kung–Foo Whing.

Telegraph cable networks expanded worldwide. On 18650517 in Paris, France, the *International Telegraph Union* was founded. After failed attempts in 1857 and 1865, in 1866, the transatlantic telegraph cable connection became operational.

After a decade of court battles and technical improvements in 1880s, telegraph and telephone services expanded worldwide. Often, these services were combined with the classical post office mail service, forming the Post, Telegraph and Telephone (PTT) office. Numerous telegraph and/or telephone users were connected together via manual switch boards staffed with tens to hundreds switch boards operators. On 18780128 in New Haven CT USA *BTC* started the operation of the first manual switch boards serving 21 local telephone customers.

(1857, Heinrich Rudolf Hertz, 1894) German scientist, physicist and professor at Kiel, Karlsruhe, and Bonn was, between 1878 and 1880 at Berlin University. He was one of the ablest students of Kirchhoff and (1821, Hermann Ludwig Ferdinand von Helmholtz, 1894). In 1888 at the University of Karlsruhe, Hertz performed a number of epoch making experiments, proving existence of electromagnetic waves and validity of ElectroMagnetics-EmpiricalLaws. Hertz, and Heaviside, independently, wrote "Maxwell equations." Hertz constructed a number of different aerials/antennas, transmitters, receivers, and radars to achieve these goals. Ironically, the conclusions made in Hertz's laboratory notes, diary, and letters appear to be different from conclusions made in his published papers. A new wireless radio era began.

On 18930301 in St. Louis, United States, (1856, Nikola Tesla, 1943), Croatia born and raised, student in Graz Habsburg Empire, prolific inventor, scientist, and the sparkplug of American industrial (r)evolution, as a principal of the *Tesla Electric Company*, made a public demonstration of wireless radio communications; he used a self made spark transmitter and a Geissler like tube receiver. On 18950313∼0230, *Tesla Electric Company* Laboratories at 33–35 South Fifth Avenue in New York City were engulfed in flames, allegedly, but unlikely, set by his arch enemy (1847, Thomas Alva Edison, 1931); numerous vacuum tubes and radio equipment were destroyed. Despite being on the opposite sides in numerous public showcases, both Edison and Tesla respected each other, greatly. In the summer of 1897, Tesla established a wireless communication link between a boat on Hudson River and his laboratory in Manhattan, tens of kilometers away. In 1898, he made a demonstration in the New York City Madison Square Garden of a radio remotely controlled boat. Tesla's commercial adventures in radio failed. In 1943, the U.S. Administration declared Tesla to have the priority in the U.S. radio patent to that of Marconi.

On 18950507, (1859, Aleksandr Stepanović Popov, 1906), Professor in Petro(v)grad (St. Petersburg), Russia, using self made equipment, sent a number of radio telegraph messages and demonstrated his equipment to the Russian Navy. He gradually improved the equipment and was able, in 1900, to send messages from a Russian Navy ship in Baltic Sea to his laboratory in Petro(v)grad, about 50 km away. His successful experiments were classified as top secret. However, made in Russia industrial production was not successful; during Russian–Japanese war in 1905, Russians used German *Telefunken* made radio equipment.

Japanese efforts in radio telegraphy were started in 1897 by Mitsuru Saheki of *ETL*, who was, by 1900, able to reach distances of 50 km. By 190310 a 1200-km-long radio link between Nagasaki Japan and Taiwan was established. By 1904, radio communications over distances of 80 nm (150 km) were routinely made. Made in Japan radio sets, a joint effort by professor Shunkichi Kimura of *Sendai High School* (now *Tohoku University*) and Matsunosuke Matsuhiro of *Ministry of Posts and Telecommunications*, were installed on the ships of Japanese Navy, victors of 19050527 Battle of Tsushima in the Sea of Japan. In 192603, Instructor Shintano Uda at *Tohoku University* in Sendai invented and designed a new antenna; his boss Professor Hidetsugu Yagi described the invention in English, thus Yagi–Uda antenna.

In 189602, (1874, Guglielmo William 1914 Sir Senator 1929 Marchese Marconi, 1937), a pupil of professor (1850, Augusto Righi, 1920), but without any formal education, first a system integrator and user of professor (1844, Eugène Édouard Désiré Branly, 1940) coherer and soon a master of antenna and wireless radio equipment design, arrived from his native Italy to his Irish mother country, the British Empire. With the help and extensive connections of his grandfather, whiskey magnate, Andrew Jameson, young Marconi demonstrated promptly his wireless radio equipment to the British establishment. In 1897, he cofounded *The Wireless Telegraph and Signal Company Limited*, which in 1900 became the *Marconi's Wireless Telegraph Company*. In the summer of 1898, Marconi established the first wireless ship-to-shore telegraph link with the British Royal yacht HMS Osborne. In 190112, Marconi must had been a very desperate man to undertake an experiment under difficult wintery blustering gale conditions near St. Johns, Newfoundland, then British Empire, now Signal Hill Canada. On 19011212 he announced a successful reception of radio signals from Poldhu England transmitter—a highly improbable event, which was neither monitored nor confirmed by any independent source. Marconi became a successful businessman, ladies man, and a true ambassador of radio industry and science. The Marconi Company lead the development of numerous radio networks, ship communications, and the British Imperial Wireless Telegraph Chain radio communications network.

In 1906, during British Marconi Corporation against German Telefunken wireless telegraphy commercial wars, the First International Radiotelegraph Conference was held in Berlin, German Empire; the Conference established the *International Radiotelegraph Convention*; the annex to this Convention contained the first regulations governing wireless > radio telegraphy; this annex has been subsequently revised and expanded by following radio conferences, and emerged into a text now known as the *Radio Regulations.*

The first wireless telegraph radio bubble, that is, the stock scam, started far too soon; in 191112 (1869, Arthur Frederick Collins, 195x), a radio expert and entrepreneur, and four officers of the Continental Company were indicted for the purpose of selling worthless stock in the Collins Wireless Telephone Company; in the same year, (1873, Lee De Forest, 1961)'s Radio Telephone Company and the United Wireless Company both went bankrupt; in 1912, the later company was acquired by the British Marconi Company and sold to American Marconi Company; in 1912, De Forest was arrested but cleared of all charges in 1913; however, in the same year 1913 A.F. Collins and accomplices were sentenced and jailed. Brits were not immune to fraudulent behavior and political corruption, either; on the contrary, in what has been known as the Marconi Scandal, the Director of the British Marconi Company, his brother the Attorney General, their friend the Postmaster General, the Chancellor of the Exchequer, some other ministers, and smaller fish were caught, red handed, in, to put it mildly, insiders trading; some of these Honorable men became knights, barons, High Commissioners, Prime Ministers, and so on, while some men of honor, who opposed their scam, were sent to killing fields of Somme and Verdun, many never to return.

On 19120414 around 2340, the *unsinkable* Titanic was perforated by an iceberg; two hours later, on 19120415 around 0140 the majestic ship capsized; the heroics of Marconi Company wireless operators saved 711/2201 lives; the same day, the stock of American Marconi Company jumped from \$55 to \$225. In 1919, the American Marconi Company was forced to sell itself to the American owners AT&T and General Electric (GE), who formed the Radio Corporation of America (RCA); some privileged ones bought a stock for a few dollars. In 1927, the stock was worth \$100/5; in 1929 it peaked at \$573/5; on 19291029, the *butcher of Gallipoli* visited the Wall Street; the Dow Jones Industrial Average plunged about 13%, an event known as the *Black Tuesday*; the RCA stock dipped at \$15/5.

In 1933, during the Great Recession, A.F. Collins' brother M.H. Collins and M.H.'s son (1909, Arthur Andrew Collins, 1987) established the Collins Radio Company.

On 19150523, Italy treacherously broke her treaty with German and Habsburg Empires, declared war on them; the same night, Italian armies crossed the mutual border; the twelve *Battles of Isonzo Soča*, the most bloody and costly battles of the World War I, claimed the life of (1874, Friedrich Fritz Hasenöhrl, 1915), a radio expert, former student and successor to (1844, Ludwig Eduard Boltzmann, 1906) who himself perished in nearby Duino Castle; Hasenöhrl's pupil (1887, Erwin Rudolf Josef Alexander Schrödinger, 1961) distinguished himself as an artillery officer in the Austro-Hungarian 5th Army of (1856, Svetozar Boroević 1905 von Bojna 1918 Feldmarshal, 1920), one of the most successful and the most decorated Field Marshal of the World War I; (1886, Oskar Kokoschka, 1980), expressionist painter and writer, companion of dashing composer widow (1879, Alma Schindler 1901 Mahler 1911 Gropius 1929 Welfer, 1964), and decorated severely wounded veteran of Galizia Front, was also at Soča. During the famous *Battle of Kobarid* (*Karfreit, Caporetto*), also known as the *12th Battle of Isonzo Soča*, from 19171024 to 19171109, Marconi was a part of Italian contingent that escaped the rout of the 400,000 men weak Italian Army by combined German and so-called Austro-Hungarian forces; ironically, three-fifth of the Habsburg Empire (called Austria–Hungary by western misinformation spin doctors) composed of Slavic population, and at Soča in particular, the bravest and most decorated soldiers were Croats and Bosniacs; massive retreat of Italian armies was followed by the change of Italian military and political structure; among numerous casualties of *Battles of Isonzo Soča* were two younger brothers of mathematician (1875, Giuseppe Beppo Levi, 1961), engineer (1885, Decio Levi, 1917) and mathematician (1883, Eugenio Elia Levi, 1917); Beppo Levi, as a dwarf, escaped the World War I, but emigrated to Argentina in 1938. Briefly in 191806/07 on Piave front, one of ambulance drivers on Italian side was (1899, Ernest Miller Hemingway, 1961) who described some of the action in his novel *A Farewell to Arms.* The Grand Admiral of Habsburg's Kriegsmarine was (1851, Anton von Haus, 1917), the grandfather of (1925, Hermann Anton Haus, 2003), distinguished IEEE fellow; officers of the Habsburg Air Force included (1881, Todor Theodore von Sköllöskislaki Kármán, 1963), head of research in the Army Aviation Corps, and (1883, Richard von Mises, 1953), a pilot and aircraft constructor; later, both men became famous pioneers of aeronautical engineering and fluid mechanics and emigrated to the United States.

On 19190323 (1883, Benito Amilcare Andrea Mussolini 1922 Il Duce 1943, 1945), a well-educated reporter and severely wounded veteran of Isonzo, formed the *Fascist Party* and on 19221031 became the Italian Prime Minister and Dictator Il Duce; his partner to be (1889, Adolf Hitler Schicklgruber Rothschild? 1933 Reichskanzler 1934 Führer 1945, 1945) distinguished himself on the western fronts in both the World Wars. For his geniality, boldness, bravery, and legendary victories around Mount Matajur, which contributed significantly toward the Caporetto rout, (1891, Erwin Johannes Eugen Rommel, 1944) earned numerous ordains becoming the youngest recipient of Germany's highest medal, the *Pour le Mérite*; in the World War II, *Afrika Korps* under bold leadership of Field Marshal Rommel, nicknamed (*Wüstenfuchs, The Desert Fox*), greatly admired by his soldiers and opponents alike, although always outnumbered by 1:2–1:16, scored many victories before the final defeat and demise of *Afrika Korps* in 1943.

In 1932, Marconi established the 35-km-long 525 MHz microwave radiotelephone link between the Vatican City and the Pope's summer residence at Castel Gandolfo. The service was inaugurated in 193302. Marconi, 1909 Nobel Laureate, a distinguished member of the *Fascist Party* since 1923, since 1930 president of the Royal Academy of Italy, Member of the Fascist Grand Council, served Mussolini and Italian government well.

In 1931 (1894, Andre G. Clavier, 1972) of *International Telegraph and Telephone Corporation* (*ITT*) and his team demonstrated 1667 MHz microwave radiotelephone link between Calais France and Dover England. In 1934, a number of commercial service links in France and England was inaugurated. In 1941, Clavier established a number of beyond-the-horizon radio links for German Military. In 1945, Clavier emigrated to the United States.

In 1932 in Madrid, Spain, by decision of the international conference, the *International Telecommunications Union* (ITU) was formed by the merger of the *International Telegraph Union* and the *International Radiotelegraph Convention.*

In 1930s, Soviet and British scientists developed independently different radar systems; World War II, and other military conflicts, facilitated development of variety of mobile radio systems based on analog AM, FM and PM schemes; soon the first spread spectrum systems were introduced.

The era of satellite communications and navigation started on 19571004, when Soviets launched Sputnik 1, the world's first artificial satellite. Satellites that followed were primarily used for military communications, navigation, and information collection. Commercial mobile satellite systems (MSS), developed in the 1970s and early 1980s, used ultra high frequency (UHF) bands around 400 MHz and around 1.5 GHz for communications and navigation services. Modern day MSS use, mostly, the frequency bands below 3 GHz (see Figure 2.1), while some employ transponders in frequency bands 6/4, 30/20, and 44/20 GHz.

In the 1950s and 1960s, numerous private mobile radio networks, citizen band (CB) mobile radio, ham operator mobile radio, and portable home radio telephones used diverse types and brands of radio equipment and chunks of airwaves located anywhere in the frequency band from near 30 MHz to 3 GHz. Then, in 1970s, Ericsson introduced the *Nordic Mobile Telephone* (NMT) system, and AT&T Bell Laboratories introduced *Advanced Mobile Phone Service* (AMPS). The impact of these two *public land mobile telecommunication systems* on the standardization and prospects of mobile radio communications may be compared with the impact of Apple and IBM on the personal computer industry. In Europe, systems like AMPS competed with NMT systems; in the rest of the world, AMPS, backed by Bell Laboratories' reputation for technical excellence and the clout of AT&T, became *de facto* and *de jure* the technical standard (British TACS and Japanese MCS-L1 are based on). In 1982, the *Conference of European Postal and*

Telecommunications Administrations (CEPT) established Groupe Spécial Mobile (GSM) with the mandate to define future Pan-European cellular radio standards.

On January 1, 1984, during the phase of explosive growth of AMPS and similar cellular mobile radio communications systems and services came the divestiture (breakup) of AT&T. A colossal turmoil called deregulation, which still reigns almost worldwide but with a few exceptions, is using a number of different systems, all ironically called standard: boom times for predators (thieves, politicians, lawyers, bankers, wallstreeters, used car salesman, and other non-contributors), but a bust for pray (a pink slip nightmare for thousands of engineers and diligent workers throughout the industries and sectors). In such a Wild West environment, many systems have been aborted, many have been stillborn, many have died, and have been dying, in infancy; only a few systems survived, and have been surviving and/or are going into bankruptcy, but only some of the survivors would have been able to mature and serve the customers well. Some of the frequency bands used by these systems are illustrated in Figure 2.1.

2.4 Present and Future Trends

On the basis of the solid foundation established in 1970s the build up of mobile radio systems and services is continuing at an annual rate higher than 20%, worldwide. Terrestrial mobile radio systems offer analog and digital voice and low to medium rate data services compatible with existing public switching telephone networks in scope, but with poorer voice quality and lower data throughput. Wireless mobile radio data networks are expected to offer data rates as high as few Mbit/s in the near future and tens of Mbit/s in the portable environment.

Equipment miniaturization and price are important constraints on the systems providing these services. In the early 1950s a mobile radio equipment used a considerable amount of a car's trunk space and challenged the capacity of car's alternator/battery source, while in transmit mode; today, the pocket-size $\approx$100 g handheld cellular radio telephone, manual and battery charger excluded, provides a few hours of talk capacity and dozens of hours in the stand-by mode. The average cost of the least expensive models of battery powered cellular mobile radio telephones has dropped proportionally, and has broken the US $100 barrier. However, one should take the price and growth numbers with the grain of salt, since some prices and growth itself might be subsidized; many customers appear to have more than one telephone, at least during the promotion period, while they cannot use more than one telephone at the same time; these facts need to be taken into consideration while estimating growth, capacity, and efficiency of recent wireless mobile radio systems.

Mobile satellite systems are expanding in many directions: large and powerful single unit geostationary systems; medium-sized, low orbit multisatellite systems; and small-sized, low orbit multisatellite systems, launched from a plane, see Kucar (1992) and Del Re (1995). Recently, some financial uncertainties experienced by a few technically advanced LEO satellite systems, operational and planned, slowed down explosive growth in this area. Satellite mobile radio systems currently offer analog and digital voice, low to medium rate data, radio determination, and global distress safety services for travellers.

During the past 5 years numerous new digital radio systems for mobile users have been deployed. At present, users in many countries have been offered between 5 and 10 different mobile radio communications systems to choose from. There already exist radio units capable of operating on two or more different systems using the same frequency band or even using a few different frequency bands—like having a car with three different engines fuelled by gasoline, diesel, and natural gas, respectively—not the most optimal solution, indeed. Overviews of mobile radio communications systems and related technical issues can be found in [Davis, 1984], [Cox, 1987], [Mahmoud, 1989], [Kucar, 1991], [Rhee, 1991], [Steele, 1992], [Chuang, 1993], [Cox, 1995], [Kucar and Uddenfeldt, 1998], [Cimini, 1999], [Mitola, 1999], [Ariy, 1999], [Oppermann, 1999], [Akyldiz, 2001], [Cherubini, 2001], [Siegel, 2001], [Jiangzhou, 2001], [Greenstein, 2002], [Shafi, 2003] and on the web sites listed in the References section. At present, the GSM, and associated systems, spearheaded by Ericsson of Sweden, command about two-thirds of the market, worldwide; their systems, based on TDM and CDM schemes, appear to converge toward

2.5G > 2.75G > 3G > 3.5G > 4G systems carrying a number of brand names, worldwide. Their main competitor, Qualcomm of United States, promotes and sells CDMA systems based on its own patents and chips; in addition, numerous groups are trying to compete with Bluetooth of Ericsson in the market for very short range devices, and expand into the last mile market with tens of Mb/s digital radio systems. In the cellular radio market, a third team is trying to develop the China invented and sponsored Time Division Synchronous Code Division Multiple Access (TD-SCDMA) system.

2.5 Repertoire of Systems and Services for Travellers

The variety of services offered to travellers essentially consists of information in analog and/or digital form. Although most of first generation systems traffic consisted of analog or digital voice transmitted by analog frequency modulation FM (or phase modulation PM), or digital quadrature amplitude modulation (QAM) schemes, digital signaling is predominant on today's systems. By using a powerful and affordable microprocessor and digital signal processing chips, a myriad of different services particularly well suited to the needs of people on the move has been realized economically. A brief description of a few elementary systems/services currently available to travelers will follow. Some of these elementary services can be combined within the mobile radio units for a marginal increase in the cost and complexity with respect to the cost of a single service system; for example, a mobile radio communications system can include a positioning receiver, digital map, web browser, and so on.

Terrestrial systems. In a terrestrial mobile radio network labeled Mobiles T in Figure 2.2, a repeater was usually located at the nearest summit offering maximum service area coverage. As the number of users increased, the available frequency spectrum became unable to handle the increased traffic, and a need for frequency reuse arose. The service area was split into many small subareas called cells, and the term *cellular radio* was born. The frequency reuse offers an increased overall system capacity, while the smaller cell size can offer an increased service quality, but at the expense of increased complexity of the user's terminal and network infrastructure. The trade-offs between real estate availability (base stations) and cost, the price of equipment (base and mobile), network complexity, and implementation dynamics dictate the shape and the size of cellular network. Increase in the overall capacity calls for new frequency spectra, smaller cells, which requires reconfiguration of existing base stations locations; this is usually not possible in many circumstances, which leads to suboptimal solutions and even less efficient use of frequency spectrum.

Satellite systems employ one or more satellites to serve as base station(s) and/or repeater(s) in a mobile radio network. The position of satellites relative to the service area is of crucial importance for the coverage, service quality, price, and complexity of the overall network. When a satellite encompasses the Earth in 12 h, 24 h, and so on periods, the term *geosynchronous orbit* has been used. An orbit which is inclined with respect to the equatorial plane is called *inclined orbit*; an orbit with an inclination of about 90° is called *polar orbit*, (see Figure 2.3). A circular geosynchronous 24 h orbit in the equatorial plane (0° inclination) is known as the *geostationary orbit* (GSO), since from any point at the surface of the Earth the satellite appears to be stationary above the horizon; this orbit is particularly suitable for the land mobile services at low latitudes, and for maritime and aeronautical services at latitudes of $<|80|°$. Systems that use geostationary satellites include INMARSAT, AUSSAT > Optus and Thuraya. An elliptical geosynchronous orbit with the inclination angle of 63.4° is known as *Tundra orbit*. An elliptical 12 h orbit with the inclination angle of 63.4° is known as *Molniya orbit*. Both Tundra and Molniya systems have been designed to provide the coverage of the high Northern latitudes and the area around the North Pole—for users at those latitudes the satellites appear to wander around the zenith for a prolonged period of time. The coverage of a particular region (*regional coverage*), and the whole globe (*global coverage*), can be provided by different constellations of satellites including ones in inclined and polar orbits. For example, inclined near circular orbit constellations have been used by GPS (24–29 satellites, 55–63° inclination), Globalstar (48 satellites, 47° inclination), and Iridium (66 satellites, 90° inclination—polar orbits) system; all three systems provide the global coverage. Orbcomm system employs Pegasus launched inclined low earth orbit (LEO) satellites to provide uninterrupted coverage of the Earth below ±60° latitudes, and an intermittent, but frequent coverage over the polar regions.

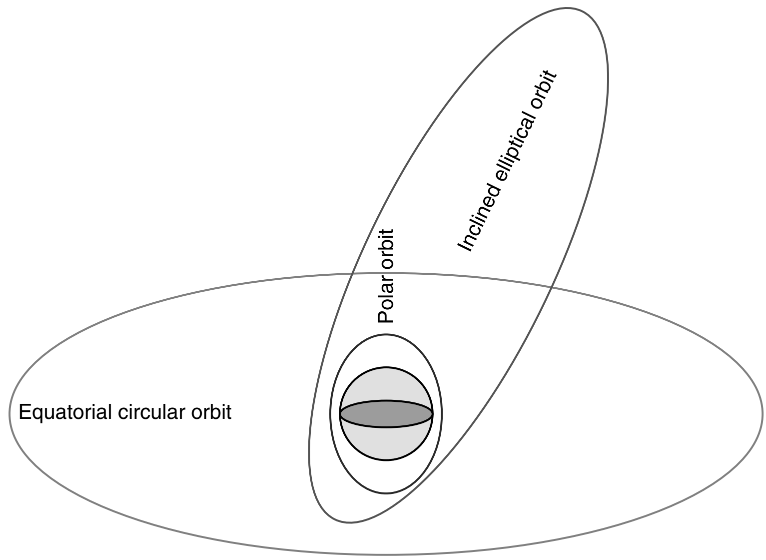


FIGURE 2.3 Equatorial circular, inclined elliptic, and polar circular orbit.

Satellite antenna system can have one (*single beam global system*) or more beams (*multibeam spot system*). The multibeam satellite system, similar to the terrestrial cellular system, employs antenna directivity to achieve better frequency reuse, at the expense of system complexity. Thuraya, one of the most advanced MSS systems, and the latest generation of INMARSAT satellites, employ more than 200 spot beams to achieve a very high frequency reuse efficiency and a coarse location as a byproduct of the spot beam system.

Radio paging is a nonspeech, one-way (from base station toward travelers), personal selective calling system with alert, without message or with defined message such as numeric or alphanumeric. A person wishing to send a message contacts a system operator by public switched telephone network (PSTN), and delivers his/her message. After an acceptable time (queuing delay), a system operator forwards the message to the traveler, by radio repeater (FM broadcasting transmitter, VHF or UHF dedicated transmitter, satellite, cellular radio system). After receiving the message, a traveler's small (roughly the size of a cigarette pack) receiver (pager) stores the message into its memory, and on demand either emits alerting tones or displays the message.

Global Distress Safety System (GDSS) geostationary GEOSAT and inclined low earth orbit satellites (LEOSAT) transfer emergency calls sent by beacons of individuals in distress to the central earth station. Examples are COSPAS, Search And Rescue Satellite Aided Tracking system (SARSAT), COSPAS–SARSAT, Geostationary Operational Environmental Satellites (GOES), and SEarch and REscue Satellite (SERES). The recommended frequency for this transmission is 406.025 MHz (Figure 2.1).

Global Positioning System (GPS), (ION, 1980, 1984, 1986, 1993). U.S. Department of Defense Navstar GPS 24–29 operating satellites in inclined orbits emit L-band (L1 = 1575.42 MHz, L2 = 1227.6 MHz, L3 = 1381.05 MHz, L4 = 1841.40 MHz, L5 = 1176.45 MHz) spread spectrum signals from which an intelligent microprocessor-based receiver extracts, among others, extremely precise time and frequency information, and accurately determine its own three-dimensional position, velocity, and acceleration, worldwide. The coarse accuracy of <100 m available to commercial users has been demonstrated by using a handheld receiver. An accuracy of meters or centimeters is possible by using the precise (military) codes and/or differential GPS (additional reference) principles and kinematic phase tracking. GPS constellation in the (Mean anomaly, Ascending node longitude) coordinates and at the time of vernal equinox of the year 2002 is shown in Figure 2.4; satellites are located in six planes each spaced about 60° apart; the mean inclination is about 54.8°.

A view of the sky and GPS satellite orbits from the North Pole is shown in Figure 2.5; there is a $(90 - 54.8283 \sim 35^\circ)$ hole around the zenith, where GPS satellites do not travel; a user antenna can see 10 GPS satellites, shown as light gray large dots, above the horizon. The horizon is represented by

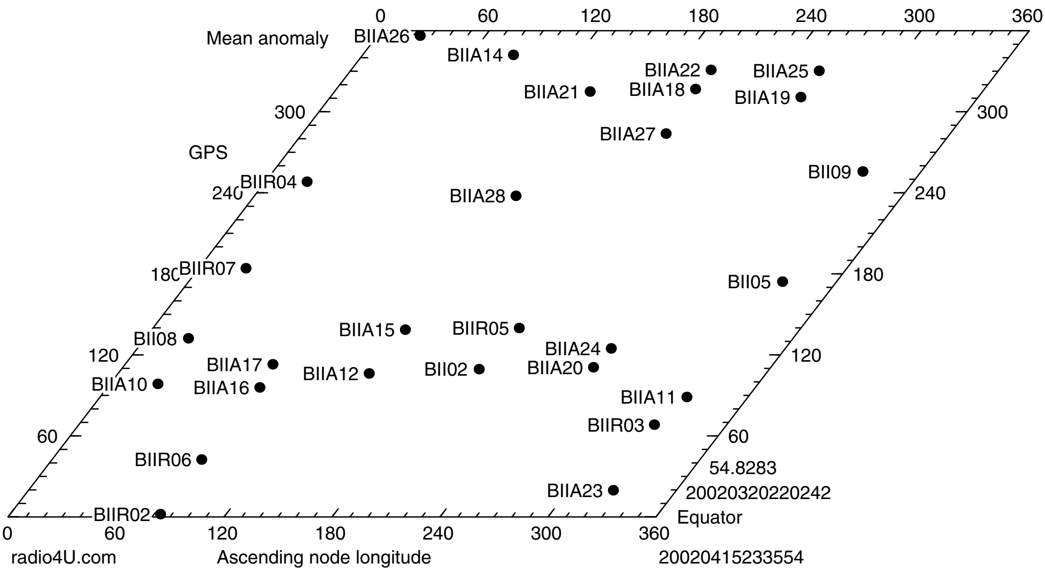


FIGURE 2.4 Global positioning system on 2002 vernal equinox: the full constellation.

the second largest concentric circle illustrated in Figure 2.5. There are additional 6 GPS satellites seen below the horizon but above $-10°$ elevation, shown as large darker gray dots positioned between the two largest concentric circles of the figure; at an elevation of $(81.4 - 90 \sim 8.6°)$ there is the GSO; there, along with the GSO, 9 INMARSAT WAAS satellites can be seen as dark gray ellipses (see also Figure 2.6); the Moon appears above the western horizon at an elevation of about 20°; the Sun, at the time of vernal equinox, is, by definition, on the horizon and, in the year 2002, at 22 h.

GLONASS (GLobal'naya NAvigatsionnaya Sputnikovaya Sistema, GLObal NAvigation Satellite System) is Russia's counterpart of the U.S.'s GPS. GLONASS operates in an FDM mode and uses frequencies between 1602 MHz and 1615 MHz, and between 1246 MHz and 1256 MHz, to achieve goals similar to GPS. For most of the time, the GLONASS system has been served by less than full 24 satellite constellation.

Other systems have been studied by the European Space Agency (Navsat), European radionavigation system proposal GALILEO, and by West Germany (Granas, Popsat, and Navcom). In 200512 GALILEO launched its first (experimental) satellite; a complete constellation of 30 satellites would achieve a radionavigation and radiolocation precision at least equaling that of GPS and GLONASS, see also Figure 2.1 and Kucar (2006b). In recent years many payloads carrying navigation transponders have been puton board of GSO satellites; corresponding navigational signals enable an increased overall availability and improved determination of user positions. The comprehensive projects, which may include existing and new radionavigation payloads, in space and at the earth, include mostly United States' *Wide Area Augmentation System* (WAAS), Japanese *Multi-functional Transport SATellite* (MTSAT), European *European Geostationary Navigation Overlay Service* (EGNOS), which seems to be a predecessor of the GALILEO project, and an international undertaking with a rather bold and complex task of integrating most of the previous systems into a compatible *Global Navigation Satellite System* (GNSS) (see Kucar, 2006b).

LORAN–C is the 100 kHz frequency navigation system, which provides a positional accuracy between 10 and 150 m. A user's receiver measures the time difference between the master station transmitter and secondary stations signals, and defines his hyperbolic line of position. North American LORAN–C coverage includes the Great Lakes, Atlantic, and Pacific Coast, with decreasing signal strength and accuracy as the user approaches the Rocky Mountains from the East. Recently, new LORAN–C stations have been augmented, worldwide. Similar radionavigation systems are the 100 kHz Decca and 10 kHz Omega.

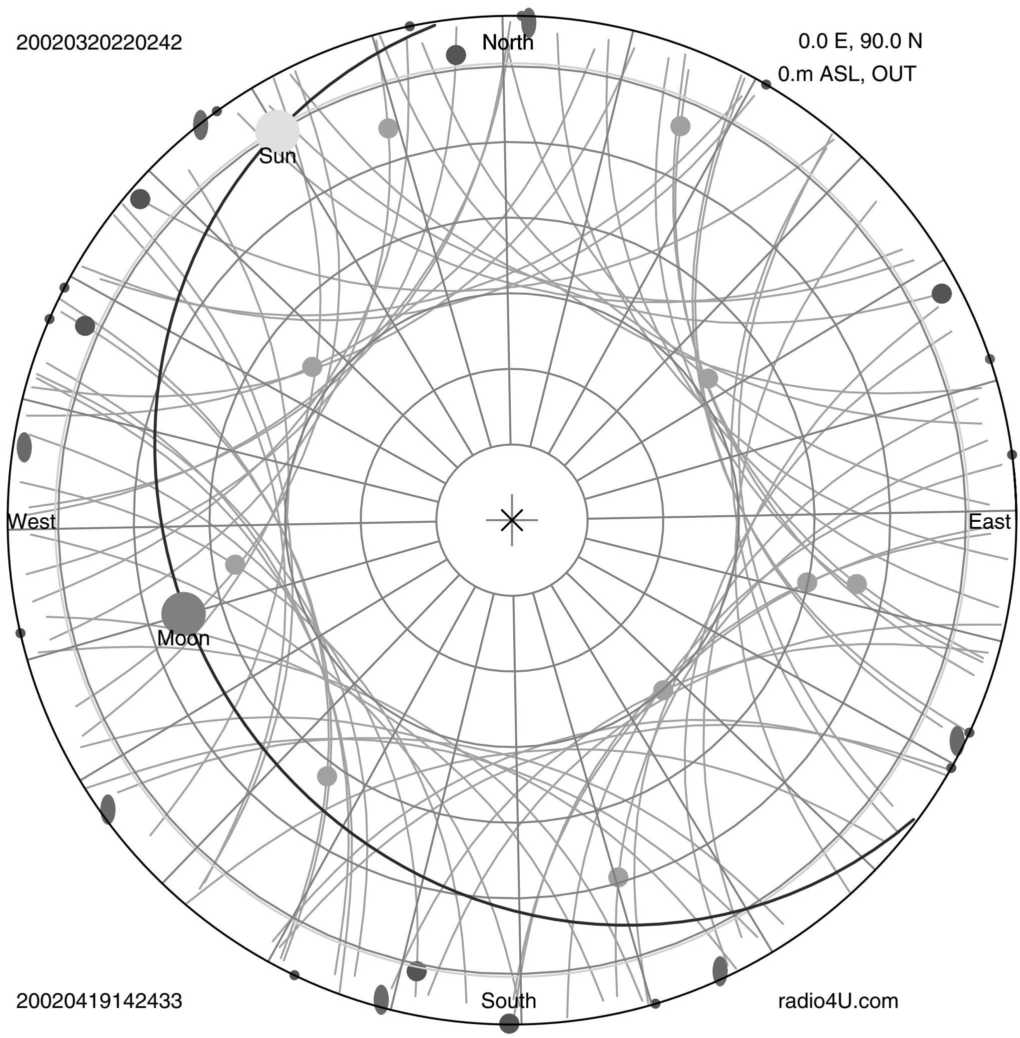


FIGURE 2.5 Global positioning system on 2002 vernal equinox: polar view.

Dispatch two-way radio land mobile or satellite system, with or without connection to the PSTN, consists of an operating center controlling the operation of a fleet of vehicles such as aircraft, taxis, police cars, tracks, rail cars, and so on. The summary of some of existing and planned terrestrial mobile radio systems, including MOBITEX RAM and ARDIS, is given in Table 2.3. OmniTRACS dispatch system employs Ku-band geostationary satellite located at 103° W to provide two-way digital message and position reporting (derived from incorporated satellite-aided LORAN–C receiver), throughout the contiguous U.S. (CONUS).

Cellular radio or public land mobile telephone system offers a full range of services to the traveler, which are similar to those provided by PSTN; however, service availability of a cellular system has been significantly lower than that of PSTN, at least the PSTN systems before recent turmoils in the telecommunications industry. The technical characteristics of some of existing and planned systems are summarized in Table 2.4.

Vehicle Information System and *Intelligent Highway Vehicle System* are synonyms for the variety of systems and services aimed toward traffic safety and location. This includes traffic management, vehicle identification, digitized map information and navigation, radionavigation, speed sensing and adaptive

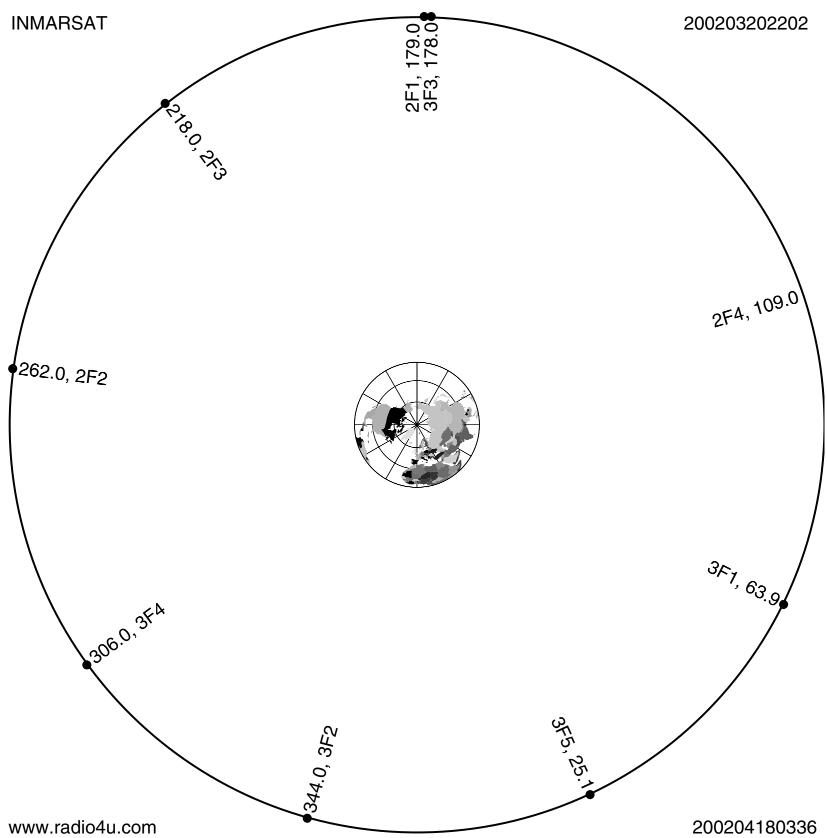


FIGURE 2.6 Inmarsat satellite in the geostationary orbit: 20020418.

TABLE 2.3 The Comparison of Dispatch WAN/LAN Systems

Parameter	United States	Sweden	Japan	Australia	CDPD	IEEE	Europe
TX frequency (MHz)						802.11	TETRA
Base	935–941	76.0–77.5	850–860	865.00–870.00	869–894	2400–2483	400s, 900s
	851–866			415.55–418.05		2470–2499	
Mobile	896–902	81.0–82.5	905–915	820.00–825.00	824–849	2400–2483	
	806–821			406.10–408.60		2470–2499	
Duplexing method	sfFDD	sFDD	sFDD	sfFDD	FDD	TDD	
Ch. spacing (kHz)	12.5	25.0	12.5	25.0	30.0	1000	
	25.0			12.5		25.0	
Channel rate (kb/s)	$\leq$9.6	1.2	1.2	$\leq$9.6	19.2	1000	
No. of traffic ch.	480	60	799	200	832	79	
	600						
Modulation type							
Voice	FM	FM	FM	FM			
Data	FSK	MSK-FM	MSK-FM	FSK	GMSK	DQPSK	$\pi/4$QPSK

sfFDD stands for semi-duplex, full duplex Frequency Division Duplex.
Similar systems are used in the Netherlands, U.K., Russia and same countries of the former USSR, and France.
ARDIS is a commercial system compatible with U.S. specs. 25 kHz spacing; 2FSK, 4FSK, $\leq$19.2 kb/s.
MOBITEX/RAM is a commercial system compatible with U.S. specs. 12.5 kHz spacing; GMSK, 8.0 kb/s.
Source: www.radio4u.com, 20061020.

TABLE 2.4 The Comparison of Cellular Mobile Radio Systems in Frequency Bands Below 1 GHz

	System Name										
Parameter	AMPS NAMPS	MCSL1 MCSL2	NMT 900	NMT 450	R. com 2000	C450	TACS UK	GSM	IS-54 IS-136	IS-95 USA	PDC Japan
TX frequency (MHz)											
Base	869–894	870–885	935–960	463–468	424.8–428	461–466	935–960	890–915	869–894	869–894	810–826
Mobile	824–849	925–940	890–915	453–458	414.8–418	451–456	890–915	935–960	824–849	824–849	890–915
B/M max eirp (dBW)	22/5	19/7	22/7	19/12	20/10	22/12	22/8	27/9	27/9	/ −7	/5
Multiple access	F	F	F	F	F	F	F	F/T	F/T	F/C	F/T
Duplex method	FDD	FDD	FDD	FDD	FDD	FDD	FDD	FDD	FDD	FDD	FDD
Channel bw (kHz)	30.0 10.0	25.0 12.5	12.5	25.0	12.5	20.0 10.0	25.0 12.5	200.0	30.0	1250	25
Channels/RF	1	1	1	1	1	1	1	8	3	42	3
Channels/band	832 2496	600 1200	1999	200	160	222 444	1000	125 × 8	832 × 3	n × 42	640 × 3
Voice/Traffic:	Analog	Analog	Analog	Analog	Analog	Analog	Analog	RELP	VSELP	CELP	VSELP
comp. or kb/s	2:1	2:1	2:1	2:1	2:1	2:1	2:1	13.0	8.0	≤9.6	6.7
modulation	PM	PM	PM	PM	PM	PM	PM	GMSK	$\pi/4$	B/OQ	$\pi/4$
kHz and/or kb/s	±12	±5	±5	±5	±2.5	±4	±9.5	270.833	48.6	1228.8	42.0
Control:	Digital	Digital	Digital	Digital	Digital	Digital	Digital	Digital	Digital	Digital	Digital
modulation	FSK	FSK	FFSK	FFSK	FFSK	FSK	FSK	GMSK	$\pi/4$	B/OQ	$\pi/4$
bb waveform	Manch.	Manch.	NRZ	NRZ	NRZ	NRZ	Manch.	NRZ	NRZ	NRZ	NRZ
kHz and/or kb/s	±8.0/10	±4.5/0.3	±3.5/1.2	±3.5/1.2	±1.7/1.2	±2.5/5.3	±6.4/8.0	270.833	48.6	1228.8	42.0
Channel coding:	BCH	BCH	B1 Hag.	B1 Hag.	Hag.	BCH	BCH	RS	Conv.	Conv.	Conv.
base → mobile	(40,28)	(43,31)	burst	burst	(19,6)	(15,7)	(40,28)	(12,8)	1/2	6/11	9/17
mobile → base	(48,36)	a.(43,31) p.(11,07)	burst	burst	(19,6)	(15,7)	(48,36)	(12,8)	1/2	1/3	9/17

Multiple access: F = frequency division multiple access (FDMA); F/T = hybrid frequency/time DMA; F/C = hybrid frequency/code DMA. $\pi/4$ corresponds to the $\pi/4$ shifted differentially encoded QPSK with $\alpha = 0.35$ square root raised-cosine filter for IS-136 and $\alpha = 0.5$ for PDC. B/OQ corresponds to the BPSK outbound and OQPSK modulation scheme inbound. comp. or kb/s stands for syllabic compandor or speech rate in kb/s; kHz and/or kb/s stands for peak deviation in kHz and/or channel rate kb/s. IS-634 standard interface supports AMPS, NAMPS, TDMA and CDMA capabilities. IS-651 standard interface supports **A** GSM capabilities and **A**+ CDMA capabilities.

Source: www.radio4u.com, 20061020.

cruise control, collision warning and prevention, and so on. Some of the vehicle information systems can easily be incorporated in mobile radio communications transceivers to enhance the service quality and capacity of respective communications systems.

Infrared Data Association (IrDA) is also a synonym for short, centimeter, and up to 10-m range, but line-of-sight communications with data rates exceeding 100 Mb/s; examples include car key remote openers, TV commanders, computer IrDA ports, and so on. Competing systems, which use ISM radio frequency bands, are ZigBee, Bluetooth, and many others.

2.6 The Airwaves Management

The airwaves (frequency spectrum and the space surrounding us) are a limited natural resource shared among several different radio users (military, government, commercial, public, amateur, etc.). Its sharing (among different users, services described in the previous section, TV and sound broadcasting, etc.), coordination, and administration is an ongoing process exercised on national, as well as on international levels. National administrations such as Federal Communications Commission (FCC) in the United States, Department of Communications (DOC), now Industry Canada, in Canada, and so on, and multinational administration such as European Telecommunications Standards Institute (ETSI), in cooperation with users and industry, set the rules and procedures for planning and utilization of scarce frequency bands. These plans and utilizations have to be further coordinated internationally, bilaterally and at ITU.

Since 1947, ITU has been a specialized agency of the United Nations, stationed in Geneva, Switzerland, with more than 150 government and corporate members, responsible for all policies related to Radio, Telegraph, and Telephone. According to the ITU, the world is divided into three regions:

- Region 1—Europe (excluding Greenland) but including former Soviet Union, Outer Mongolia, Africa, and the Middle East west of Iran
- Region 2—the Americas, and Greenland
- Region 3—Asia (excluding parts west of Iran and Outer Mongolia), Australia, and Oceania

Historically, these three regions have developed, more or less independently, their own frequency plans, which best suit local purposes. With the advent of satellite services and globalization trends, the coordination between different regions becomes more urgent. Frequency spectrum planning and coordination is performed through ITU's bodies such as: Comité Consultatif de International Radio (CCIR), now ITU-R, International Frequency Registration Board (IFRB), now ITU-R, World Administrative Radio Conference (WARC), Regional Administrative Radio Conference (RARC), and so on. ITU's Study Groups (SG) particularly relevant to this contribution include

1. SG 1—Spectrum management
2. SG 3—Radiowave propagation
3. SG 4—Fixed-satellite service
4. SG 6—Broadcasting services
5. SG 7—Science services
6. SG 8—Mobile, radiodetermination, amateur and related satellite services
   a. Working Party 8A—Land mobile service excluding IMT-2000; amateur and amateur-satellite service
   b. Working Party 8B—Maritime mobile service including Global Maritime Distress and Safety System (GMDSS); aeronautical mobile service and radiodetermination service
   c. Working Party 8D—All mobile satellite services and radiodetermination satellite service
   d. Working Party 8F—IMT-2000 and systems beyond IMT-2000
7. SG 9—Fixed service

ITU-R, through its Study Groups, deals with technical and operational aspects of radio communications. Results of these activities have been summarized in the form of Reports and Recommendations published every 4 years, or more often (ITU, 1990). IFRB, now ITU-R, serves as a *custodian* of common and scarce natural resource—the *airwaves*; in its capacity, the IFRB records radio frequencies, advises the members on technical issues, and contributes on other technical matters.

On the basis of the work of ITU-R and the national administrations, ITU Members convene at appropriate RARC and WARC meetings, where documents on frequency planning and utilization, the *Radio Regulations*, are updated. The actions on a national level follow (see RadioRegs, 2004; WARC, 1992; WRC, 1997). For updated regulations and recommendations consult the ITU web site; link and address are provided at page 2-40.

The far-reaching impact of the mobile radio communications on economies and the well being of the three main trading blocks, other developing and third world countries, potential manufacturers and users, makes the airways (frequency spectrum) even more important.

ITU recommends the composite bandwidth-space-time domain concept as a measure of spectrum utilization. The spectrum utilization factor $U = B \cdot S \cdot T$ is defined as a product of the frequency bandwidth B, spatial component S, and time component T. As mobile radio communications systems employ simple omnidirectional antennas, their S factor will be rather low; since they operate in a single channel arrangement, their B factor will also be low. New digital schemes tend to operate in packet/block switching modes that are inherently loaded with a significant amount of overhead and idle traffic, and their T factor will be low as well—consequently, mobile radio communications systems will have, in comparison with *point-to-point* (P2P) fixed radio systems, a low to poor spectrum utilization factor.

The model of a mobile radio environment, which may include different sharing scenarios with fixed service and other radio systems, can be described as follows; Objects of our concern are events (e.g., conversation using a mobile radio, measurements of amplitude, phase and polarization at the receiver) occurring in time $\{u^0\}$, space $\{u^1, u^2, u^3\}$, spacetime $\{u^0, u^1, u^2, u^3\}$, frequency $\{u^4\}$, polarization $\{u^5, u^6\}$, and airwaves $\{u^0, u^1, u^2, u^3, u^4, u^5, u^6\}$ (see Table 2.5). The coordinate $\{u^4\}$ represents frequency resource, that is, bandwidth in the spacetime $\{u^0, u^1, u^2, u^3\}$. Our goal is to use a scarce natural resource—the airwaves in an environmentally friendly manner.

When users/events are divided (sorted, discriminated) along the time coordinate u^0, the term time division is employed for function $f(u^0)$. A division $f(u^4)$ along the frequency co-ordinate u^4 corresponds to the frequency division. A division $f(u^0, u^4)$ along the coordinates (u^0, u^4) is usually called a code division or frequency hopping. A division $f(u^1, u^2, x^3)$ along the coordinates (u^1, u^2, u^3) is called the space division. Terrestrial cellular and multibeam satellite radio systems are vivid examples of the space division concepts. Coordinates $\{u^5, u^6\}$ may represent two orthogonal polarization components, horizontal and vertical or right-handed and left-handed circular; a division of users/events according to their polarization components may be called the polarization division. Any division $f(u^0, u^1, u^2, u^3, u^4, u^5, u^6)$ along the

TABLE 2.5 The Multidimensional Spaces Including the Airwaves

<table>
<tr><td>u^0</td><td>Time</td><td rowspan="4">Spacetime</td><td rowspan="7">Airwaves</td></tr>
<tr><td>u^1</td><td rowspan="3">Space</td></tr>
<tr><td>u^2</td></tr>
<tr><td>u^3</td></tr>
<tr><td>u^4</td><td colspan="2">Frequency/bandwidth</td></tr>
<tr><td>u^5</td><td colspan="2" rowspan="2">Polarization</td></tr>
<tr><td>u^6</td></tr>
<tr><td>u^7</td><td colspan="3" rowspan="3">Doppler</td></tr>
<tr><td>u^8</td></tr>
<tr><td>u^9</td></tr>
<tr><td>u^A</td><td colspan="3">Users: government/military, commercial/public, fixed/mobile, terrestrial/satellite</td></tr>
<tr><td>u^B</td><td colspan="3" rowspan="3"></td></tr>
<tr><td>$\vdots$</td></tr>
<tr><td>u^n</td></tr>
<tr><td colspan="3">Source: www.radio4u.com, 20061020</td><td>Airwaves</td></tr>
</table>

coordinates ($u^0, u^1, u^2, u^3, u^4, u^5, u^6$) may be called the airwaves division. Coordinates $\{u^7, u^8, u^9\}$ may represent velocity (or Doppler frequency) components; a division of users/events according to their Doppler frequencies similar to the moving target indication (MTI) radars may be called the Doppler frequency division. We may also introduce coordinate $\{u^A\}$ to represent users, divide the airways along the coordinate $\{u^A\}$ (military, government, commercial, public, fixed, mobile, terrestrial, satellite, and others) and call it the users division. Generally, the segmentations of frequency spectra to different users lead to uneven use and uneven spectral efficiency factors among different segments.

In analogy with division, we may have time, space, frequency, code, airwaves, polarization, Doppler, users, $\{u^\alpha, \ldots, u^\omega\}$ access, and diversity. Generally, the signal space may be described by m coordinates $\{u^0, \ldots, u^{m-1}\}$. Let each signal component has k degrees of freedom. At transmitter site, each signal can be radiated via n_T antennas, and received at n_R receiver antennas. There is a total of $n = n_T + n_R$ antennas, two polarization components, and L multipath components, that is, paths between transmitter and receiver antennas. Thus, the total number of degrees of freedom $m = k \times n \times 2 \times L$. For example, in a typical land mobile environment there can exist $L = 16$ multipath components; if one wants to study a system with four antennas on the transmitter side and four antennas on the receiver side, and each antenna may employ both polarizations, then the number of degrees of freedom equals $16 \times 4 \times 4 \times 2 \times k = 512 \times k$. By selecting a proper coordinate system and using inherent system symmetries, one might be able to reduce the number of degrees of freedom to a manageable quantity.

2.7 Operating Environment

A general configuration of terrestrial *Fixed Service* (FS) radio systems, sharing the same space and frequency bands with *Fixed Satellite Service* (FSS) and/or *Mobile Satellite Service* MSS systems, is illustrated in Figure 2.2. The emphasis of this contribution is on mobile systems; however, it should be appreciated that mobile systems may be required to share the same frequency band with fixed systems. A *satellite system* usually consists of many earth stations, denoted Earth Station 0 . . . Earth Station 2 in Figure 2.2, one or many space segments, denoted Space Segment 0 . . . Space Segment N, and in the case of a *mobile satellite system* different types of mobile segments denoted by ♣ and ♠ in the same figure. Links between different Space Segments and mobile users of MSS systems are called *service links*; links connecting Space Segments and corresponding Earth Stations are called *feeder links*. FSS systems employ Space Segments and fixed Earth Station segments only; corresponding connections are called *service links*. Thus, technically similar connections between Space Segments and fixed Earth Station segments perform different functions in MSS and FSS systems and are referred by different names. Administratively, the feeder links of MSS systems are often referred as FSS.

Let us briefly assess spectrum requirements of an MSS system. There exist many possibilities of how and where to communicate in networks shown in Figure 2.2. Each of these possibilities can use different spatial and frequency resources, which one needs to assess for sharing purposes. For example, a mobile user ♣ transmits at frequency f_0 using a small hemispherical antenna toward the Space Segment 0. This Space Segment 0 receives a signal at frequency f_0, transposes it to frequency F_{n+0}, amplifies it and transmits it toward the Earth Station 0. This station processes the signal, makes decisions on the final destination, sends the signal back toward the same Space Segment 0 which receives the signal at frequency f_{m+0}. This signal is transposed further to the frequency F_{k+0} and emitted via inter satellite link ${}_0\mathrm{ISL}_1$ toward Space Segment 1, which receives this signal, processes it, transposes it, and emits toward the earth and mobile ♠ at frequency F_1. In this process a quintet of frequencies ($f_0, F_{n+0}, f_{m+0}, F_{k+0}, F_1$) is used in one direction. Once the mobile ♠ receives the signal, its set rings and sends back the signal in reverse directions at a different frequency quintet (or a different time slot, or a different code, or any combination of time, code, and frequency), thus establishing the two-way connection. Obviously, this type of MSS system uses significant parts of the frequency spectrum.

Mobile satellite systems consist of two directions with very distinct properties. A direction from an Earth Station, also called Hub and Base Station, which may include a Network Management Center (NMC), toward the satellite space segment and further toward a particular mobile user is known as the forward direction. In addition, we will call this direction dispatch direction, broadcast direction, or division direction, since the NMC dispatches/broadcasts data to different users and data might be divided in frequency (F), time (T), code (C), or a hybrid (H) mode. The opposite direction from a mobile user toward satellite space segment and further toward the NMC is known as the return direction. In addition, we will call this direction access direction, since mobile users usually need to make attempts to access the mobile network before a connection with NMC can be established; in some networks the NMC may poll the mobile users, instead. A connection between NMC and a mobile user, or between two mobile users, may consist of two or more hops, including inter satellite links, as shown in Figure 2.2.

While traveling, a customer—*user of cellular mobile radio system*—may experience sudden changes in signal quality caused by his movements relative to the corresponding base station and surroundings, multipath propagation and unintentional jamming such as man-made noise, adjacent channel interference, and cochannel interference inherent to the cellular systems. Such an environment belongs to the class of nonstationary random fields, of which experimental data are difficult to obtain, their behavior hard to predict and model satisfactorily. When reflected signal components become comparable in level to the attenuated direct component, and their delays comparable to the inverse of the channel bandwidth, *frequency selective fading* occurs. The reception is further degraded due to movements of a user, relative to reflection points and relay station, causing the Doppler frequency shifts. The simplified model of this environment is known as the *Doppler Multipath Rayleigh Channel.*

The existing and planned cellular mobile radio systems employ sophisticated narrowband and wideband filtering, interleaving, coding, modulation, equalization, decoding, carrier and timing recovery, and multiple access schemes. The cellular mobile radio channel involves a *dynamic interaction* of signal arrived via different paths, adjacent- and cochannel-interference, and noise. Most channels exhibit some degree of memory, which description requires higher order statistics of—*spatial and temporal*—multidimensional random vectors (amplitude, phase, multipath delay, Doppler frequenc, etc.) to be employed.

A two hop (uplink and downlink) link budget (flow of signal levels) of a (30 GHz/20 GHz) GSO satellite system is illustrated in a self explanatory (Figure 2.7).

A model of a multihop satellite system that incorporates interference and nonlinearities is illustrated and described in Figure 2.8. The signal flow in the forward/broadcast direction, from Base to Mobile User, is shown on the left-hand side of the picture; the right-hand side of the same picture corresponds to the reverse/access direction. For example, in the forward/broadcast direction, the transmitted signal at the Base, shown in the upper-left-hand side of the picture, is distorted due to nonlinearities in the RF power amplifier; this signal distortion is expressed via differential phase and differential gain coefficients DP and DG, respectively. The same signal is emitted toward the satellite space segment receiver denoted as point 2; here, noise N, interference I, delay τ, and Doppler frequency $_D$f symbolize the environment. The signals are further processed, amplified, and distorted at stage 3, and radiated toward the receiver 4; here again, noise N, interference I, delay τ, and Doppler frequency $_D$f symbolize the environment. The signals are translated and amplified at stage 5 and radiated toward the Mobile User at stage 6; here, additional noise N, interference I, delay τ, and Doppler frequency $_D$f characterize the corresponding environment. This model is particularly suited for a detailed analysis of the link budget and for the equipment design purposes. A system provider and cell designer may use a statistical description of a mobile channel environment, instead.

An FSS radio environment is described as the *GaussChannel*; the mean value of corresponding radio signal is practically constant and its value can be predicted with a standard deviation of a fraction of a dB. A terrestrial mobile radio channel could exhibit a dynamics of about 80 dB (a strong signal is necessary to provide coverage inside buildings, elevators, shafts, etc.) and its mean signal could be predicted with a standard deviation of 5–10 dB. This may require the evaluation of usefulness of existing radio channel models and eventual development of more accurate ones. However, the capacity of modern digital cellular systems very much depend on estimation of signal levels in a cell; consequently, sophisticated signal estimation and tracking algorithms have been developed for these purposes.

Cell engineering, prediction of service area and service quality, in an ever changing mobile radio channel environment, is a very difficult task. The average path loss depends on terrain micro structure

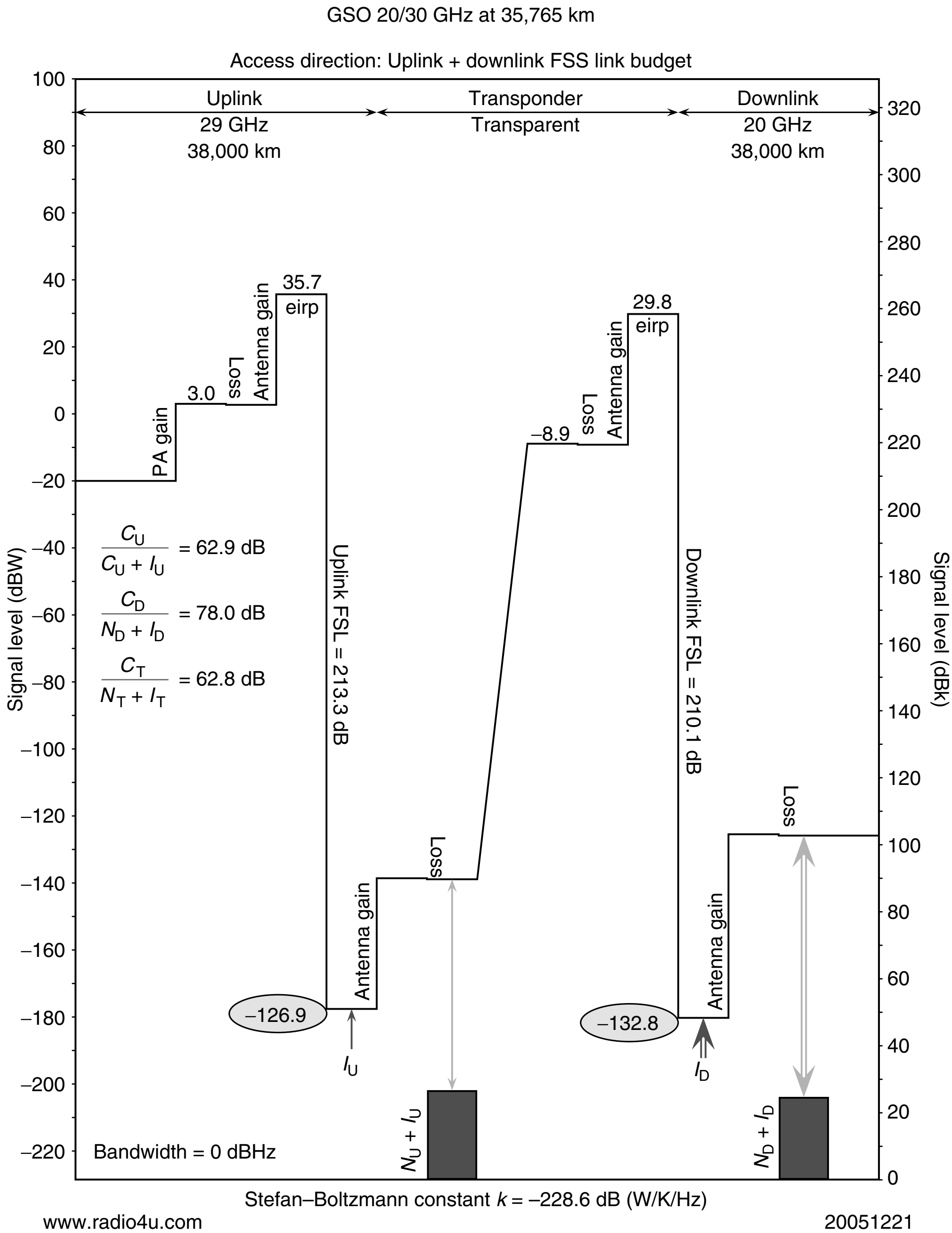
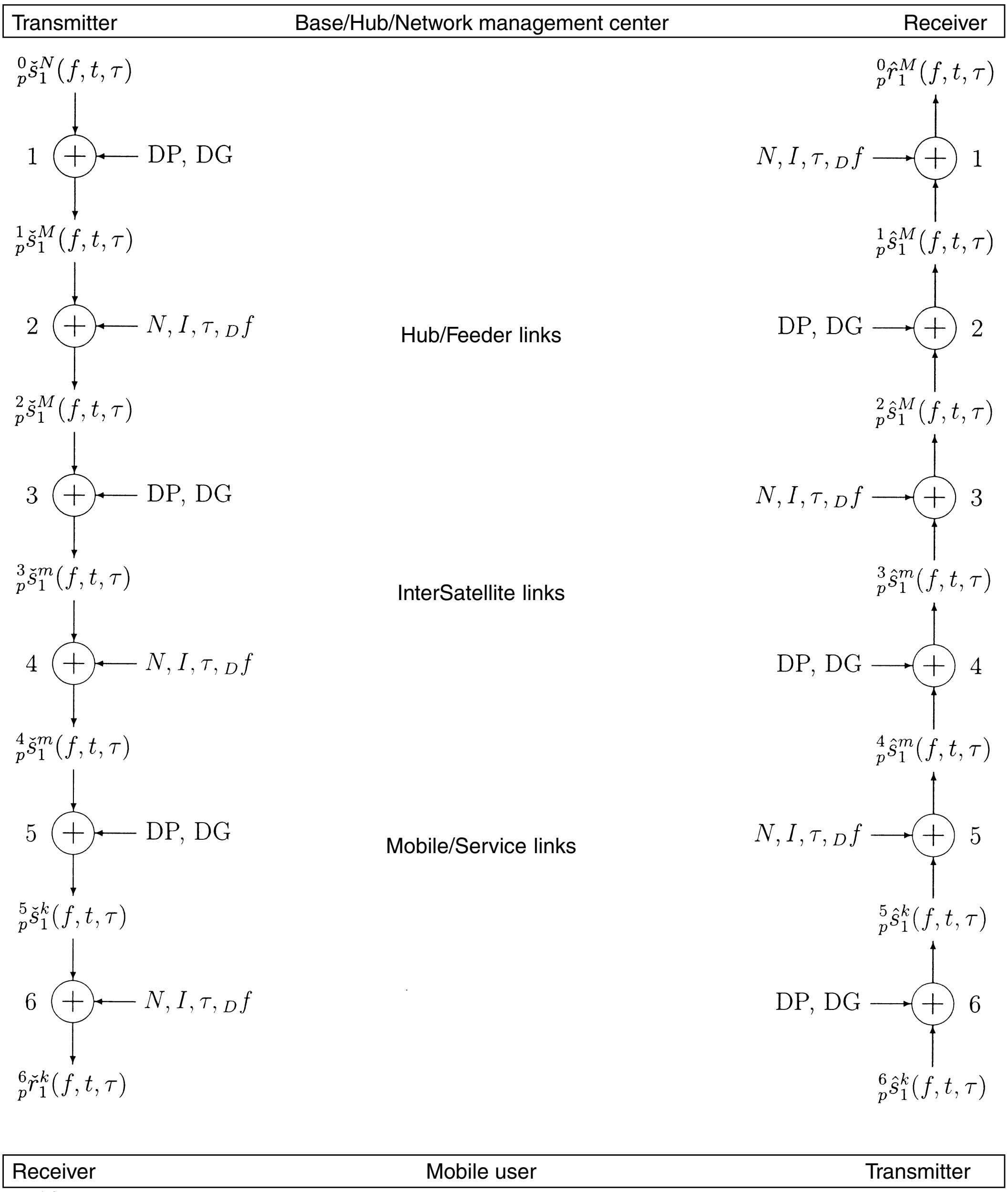


FIGURE 2.7 The link budget of a 30/20 GHz satellite system.

within a cell, with considerable variation between different types of cells, that is, urban, suburban, and rural environments. A variety of models based on experimental and theoretic work have been developed to predict path radio propagation losses in a mobile channel. Unfortunately, none of them is universally applicable. In almost all cases, an excessive transmitting power is necessary to provide an adequate system performance.

The *first generation* mobile satellite systems employ geostationary satellites (or payloads piggy-backed on a host satellite) with small 18 dBi antennas covering the whole globe. When the satellite is positioned



$^q_px^M_1(f,t,\tau)$ represents signals $x = r, s$, where r is the received and s is the sent/transmitted signal, $\breve{x}$ represents the dispatch/forward direction, and $\hat{x}$ represents access/return direction; p is the polarization of the signal at location q and the number of signal components ranges from 1 to M;

f, t, τ are frequency, time and delay of a signal at the location q, respectively.

DP and DG are Differential Phase and Differential Gain (include AM/AM and AM/PM);

$N, I, \tau, {}_Df$ are the noise, interference, absolute delay, and Doppler frequency, respectively.

FIGURE 2.8 Signals and interference in a multihop satellite system.

directly above the traveler (at zenith), a near constant signal environment, known as *GaussChannel*, is experienced. The traveler's movement relative to the satellite is negligible, that is, Doppler frequency practically equals zero. As the traveller moves—north or south, east or west—the satellite appears lower on the horizon. In addition to the direct path, many significant strength reflected components are present, resulting in a degraded performance. Frequencies of these components fluctuate due to movement of traveller relative to the reflection points and the satellite. This environment is known as the *Doppler-RiceChannel*. An inclined orbit satellite located for a prolonged period of time above 45° latitude north and 106° longitude west, could provide travelers all over the United States and Canada, including the far North, a service quality unsurpassed by either geostationary satellite or terrestrial cellular radio. Similarly, a satellite located at 45° latitude north and 15° longitude east, could provide travelers in Europe with improved service quality.

Inclined orbit satellite systems can offer a low start-up cost, a near *GaussChannel* environment, and improved service quality. Low orbit satellites, positioned closer to the service area, can provide high signal levels and short (a few milliseconds long) delays, and offer compatibility with the cellular terrestrial systems. These advantages need to be weighted against network complexity, inter satellite links, tracking facilities, and so on.

High altitude platform (HAP) systems employ balloons or aircraft flying in near circular patterns located tens of kilometers above the service area. A HAP is essentially similar to a military AWAC aircraft based system. HAP deployment cost is much lower than an equivalent satellite systems, but the operational cost might be prohibitive over the long periods of time. However, HAP system might be superior in emergency and peak load conditions.

Terrestrial mobile radio communications systems provide a signal dynamics of about 80 dB and are able to penetrate walls and similar obstacles, thus providing inside building coverage. Satellite mobile radio communications systems are power limited and provide a signal dynamics of less than 15 dB; the signal coverage is in most cases limited to the outdoors, line-of-sight and up to modestly shaded areas but not the urban canyons.

Let compare the efficiency of a *Mobile Satellite Service* (MSS) with the *Fixed Satellite Service* (FSS); both services are assumed to be using the GSO space segments. A user at the equator can see the GSO arc reaching $\pm 81°$; if satellites are spaces 2° apart along the GSO, then the same user can reach about 81 satellites simultaneously by employing 81 large antennas and reuse the same frequency 81 times. An MSS user employs a hemispherical antenna having gain of about 3 dBi; consequently, he/she can effectively use only one satellite but prevent all other satellite users to employ the same frequency. An FSS user employs a 43 dBi gain antenna that points toward a desired satellite; by using the same transmit power as an MSS user but employing larger and more expensive antenna, this FSS user can effectively transmit about 40 dB (10,000 times) wider bandwidth, if available, that is, exchange 40 dB more information. The FSS user can, by adding 3 dB more power into additional orthogonal polarization channel, reuse the same frequency band and double the capacity. Furthermore, the same FSS user can use additional antennas to reach each of 81 available satellites, thus increasing the GSO arc capacity by 80 times, in comparison with an MSS user. Consequently, the FSS is powerwise 10,000 times more efficient, and spatially, by using two polarizations, up to about 160 times more efficient than corresponding MSS. Similar comparisons can be made for terrestrial systems, where *point-to-point* (P2P) systems offer the highest power and spatial efficiency. For example, a FS system can use expensive ultra high performance antennas and reuse the same frequency at the same antenna tower many times over but pointing in different spatial directions; furthermore, FS usually employs multistate, for example, 256QAM schemes, while an MSS employs 4QAM schemes; thus, in addition to the 160 times spatial advantage, the difference in capacity equals $\log_2(256/4) = 6$ times in favor of an FS system. Mobile systems employ, at least at the beginning, omnidirectional antennas; sectorial antennas, with gains of up to 23 dBi, offer an increased reuse potential, that is, they are, in terms of spatial efficiency, in between an omnidirectional antenna and a high gain antenna for P2P communications. A terrestrial mobile user's handheld antenna has seldom more than about 0 dBi gain, while an antenna of an MSS system has a gain of up to about 6 dBi. However, a P2P system employs high gain antennas, for example, 43 dBi, on both sides. Although, in a practical environment, the relationships between just described services may differ, the convenience and smallness of today's mobile systems users terminals results in low spatial and power efficiency, which may carry a substantial economic price penalty; the real cost of a mobile system might have been subsidized by some means beyond the cost of cellular telephone and traffic charges.

2.8 Service Quality

The primary and the most important measure of service quality should be *Customer Satisfaction.* The customer's needs, both current and future, should provide guidance to a service offerer and an equipment manufacturer for both the system concept and product design stages. In the early stages of the product life, mobile radio was perceived as a necessary tool for performing important tasks; recently, mobile/personal/handheld radio devices are becoming more like status symbols and fashion. Acknowledging the importance of every single step of the complex service process and architecture, attention is limited here to a few technical merits of quality:

Guaranteed quality level is usually related to a percentage of the service area coverage, for example 97%, and for an adequate percentage of time, for example 99%. In comparison, FS P2P services exhibit 100% spatial coverage and up to 99.9999% temporal coverage for a given high quality of service. Mobile services are neither in capacity nor in quality an adequate substitute for P2P radio and/or cable services. Unfortunately, a pink slip environment and frequent (mis)management changes have disastrous effect on availability of any system.

Data service quality can be described by the average bit error rate (e.g., $\text{BER} < 10^{-5}$), packet BER ($\text{PBER} < 10^{-2}$), signal processing delay (1–10 ms), multiple access collision probability (<20%), the probability of a false call (false alarm), the probability of a missed call (miss), the probability of a lost call (synchronization loss), and so on.

Voice quality is usually expressed in terms of the mean opinion score (MOS) of subjective evaluations by service users. MOS marks are bad = 0, poor = 1, fair = 2, good = 3, and excellent = 4. MOS for PSTN voice service, pooled by leading service providers, relates the poor MOS mark to a signal-to-noise ratio (S/N) in a voice channel of $\text{S/N} \sim 35$ dB, while an excellent score corresponds to $\text{S/N} > 45$ dB. Currently, the users of the mobile radio services are giving poor marks to the voice quality associated with a $\text{S/N} \approx 15$ dB and an excellent mark for $\text{S/N} > 25$ dB. It is evident that there is a significant difference (20 dB) between the PSTN and mobile services. If digital speech is employed, both the speech and the speaker recognition have to be assessed. For more objective evaluation of speech quality under *real conditions* (with no impairments, in the presence of burst errors during fading, in the presence of random bit errors at $\text{BER} = 10^{-2}$, in the presence of Doppler frequency offsets, in the presence of truck acoustic background noise, in the presence of ignition noise, etc.), additional tests such as the diagnostic acceptability measure DAM, diagnostic rhyme test DRT, Youden square rank ordering, Sino-Graeco-Latin square tests, and so forth can be performed.

1970s: customer reported a problem; a technician/engineer was sent to fix the problem.
1990s: customer reported a problem; a team of lawyers was sent to fix the customer.
20xxs: customer does NOT report a problem anymore; he/she is afraid of being sued.

2.9 Network Issues and Cell Size

To understand ideas and technical solutions offered by existing schemes, and what might be expected from future systems, one needs to analyze the reasons for their introduction and success. Cellular mobile services are flourishing at an annual rate higher than 20%, worldwide, but not uniformly. The first generation systems such as AMPS, NMT, TACS, MCS, etc. use *frequency division multiple access* (FDMA) and digital modulation schemes for access, command and control purposes and analog phase/frequency modulation schemes for the transmission of an analog voice. Most of the network intelligence is concentrated at fix elements of the network including base stations, which seem to be well suited to the networks with a modest number of medium to large-sized cells. To satisfy the growing number of potential customers, more cells and base stations were created by the cell splitting and frequency reuse process. Technically, the shape and size of a particular cell is dictated by the base station antenna pattern and the topography of the service area. Current terrestrial cellular radio systems employ cells with 0.5–50 km radius. The maximum cell size is usually dictated by the link budget, in particular the gain of a mobile antenna and available output power as well as the noise figure, antenna gain, and output power of the base station. This situation arises in a rural environment, where the demand on capacity is low and cell splitting is not necessary. The minimum cell size is usually dictated by the need for an increase in capacity; this occurs in particular in downtown cores. Practical constraints such as real estate availability and price, and construction dynamics limit the minimum cell size to 0.5–2 km. However, in such types of dense networks, the complexity of the network and the cost of service grow exponentially with the number of base stations, while the efficiency of the first generation handover procedures became inadequate. Consequently, the second generation, all digital schemes with improved handover procedures and network supervision algorithms, which handle this increasing idle traffic more efficiently, are introduced. However, handling of the information, predominantly voice, has not been improved significantly, if at all.

In the 1980s, extensive studies of then existing AMPS and NMT based systems were performed (see Davis, 1984; Mahmoud, 1989, and so on, and the references therein). On the basis of particular service quality requirements, particular radio systems and particular cell topologies, few empirical rules have been established. Antennas with an omnidirectional pattern in a horizontal direction, but with about 10 dBi gain in vertical direction provide the frequency reuse efficiency of $N_{\text{FDMA}} = 1/12$. It was anticipated that base station antennas with similar directivity in vertical direction and 60° directivity in horizontal direction (a cell is divided into six sectors) can provide the reuse efficiency $N_{\text{FDMA}} = 1/4$, this results in a threefold increase in the system capacity; if CDMA is employed instead of FDMA, an increase in reuse efficiency $N_{\text{FDMA}} = 1/4 \rightarrow N_{\text{CDMA}} = 2/3$ may be expected. However, this thus not necessarily means that a CDMA system is more efficient than a FDMA system. The overall efficiency very much depends on spatiotemporal dynamics of a particular cell and the overall network. Ironically, the most recent CDMA schemes need three codes and a pilot tone in the reverse access direction just to perform; such a signal is essentially a multistate multilevel signal, which is inherently inefficient in any cochannel environment and requires nearly linear, thus battery DC power inefficient, power amplifier. Similar is the case with an *orthogonal frequency division multiplex* (OFDM) scheme, which is just a special case of a multicarrier system with the closest spacing possible, but with the highest peak envelope fluctuations and the highest sensitivity to numerous imperfections. However, an OFDM system inherits an advantage of a narrowband system robustness to some channel distortions.

Recognizing some of limitations of existing schemes and anticipating the market requirements, the research in *time division multiple access* (TDMA) schemes aimed at cellular mobile and DCT services, and in *code division multiple access* (CDMA) schemes aimed toward mobile satellite system, cellular and personal mobile applications, followed with introduction of nearly 10 different systems. Although employing different access schemes, TDMA (CDMA) network concepts rely on a smart mobile/portable unit which scans time slots (codes) to gain information on network behavior, free slots (codes), and so on, improving frequency reuse and handover efficiency while hopefully keeping the complexity and cost of the overall network at reasonable levels. Some of the proposed system concepts depend on low gain (0 dBi) base station antennas deployed in a license-free, uncoordinated fashion; small size cells (10–1000 m in radius) and an emitted isotropic radiated power of about 10 mW (+10 dBm) per 100 kHz have been anticipated. A frequency reuse efficiency of $N = 1/9$ to $N = 1/36$ has been projected for DCT systems. $N = 1/9$ corresponds to the highest user capacity with the lowest transmission quality, while $N = 1/36$ has the lowest user capacity with the highest transmission quality. This significantly reduced frequency reuse capability of proposed system concepts will result in significantly reduced system capacity, which need to be compensated by other means including new spectra.

In practical networks, the need for a capacity (and frequency spectrum) is distributed unevenly in space and time. In such an environment, the capacity and frequency reuse efficiency of the network may be improved by *dynamic channel allocation*, where an increase in the capacity at a particular hot spot may be traded for the decrease in the capacity in cells surrounding the hot spot, the quality of the transmission and network instability. The first generation mobile radio communications systems used omnidirectional antennas at base stations; today, three-sector 120° wide cells are typical in a heavy traffic urban environment, while entry level rural systems employ omnidirectional antennas; the most demanding environments with changing traffic patterns employ adaptive antenna solutions, instead.

To cover the same area (space) with smaller and smaller cells, one needs to employ more and more base stations. A linear increase in the number of base stations in a network usually requires an $n(n-1)=2$ increase in the number of connections between base stations, and increase in complexity of switches and network centers. These connections can be realized by fixed radio systems (providing more frequency spectra will be available for this purpose), or, more likely, by a cord (wire, cable, fiber, etc.). An increase in overall capacity is attributed to

- Increase in available bandwidth and particularly above 1 GHz, but to detriment to other services (see Figure 2.2).
- Increased use of adaptive antenna solutions which, through spatial filtering, increase capacity and quality of the service, but at significant increase in cost
- Trade-offs between service quality, vehicular versus pedestrian environments, analog versus digital voice, and so on.

The *first generation* geostationary satellite system antenna beam covers the entire Earth, that is, the cell radius equals ≈ 6789 km. The *second generation* geostationary satellites use larger multibeam antennas providing 10–20 beams (cells) with 800–1600 km radius. Low orbit satellites such as Iridium use up to 37 beams (cells) with 670 km radius. The *third generation* geostationary satellite systems will be able to use very large reflector antennas (roughly the size of a baseball stadium), and provide 80–300 beams (cells) with a cell radius of $\approx$200 km. If such a satellite is tethered to a position 400 km above the Earth, the cell size will decrease to $\approx$2 km in radius, which is comparable in size with today's small size cell in terrestrial systems. Yet, such a satellite system may have the potential to offer an improved service quality due to its near optimal location with respect to the service area. Similar to the terrestrial concepts, an increase in the number of satellites in a network will require an increase in the number of connections between satellites and/or Earth network management and satellite tracking centers, and so on. Additional factors that need to be taken into consideration include price, availability, reliability, and timeliness of the launch procedures, a few large versus many small satellites, tracking stations, and so on.

2.10 Coding and Modulation

The conceptual transmitter and receiver of a mobile system may be described as follows. The transmitter signal processor accepts analog voice and/or data and transforms (by analog and/or digital means) these signals into a form suitable for a double sided suppressed carrier amplitude modulator (also called quadrature amplitude modulator [QAM]). Both analog and digital input signals may be supported, and either analog or digital modulation may result at the transmitter output. Coding and interleaving can also be included. Very often, the processes of coding and modulation are performed jointly; we will call this joint process *codulation.* A list of typical modulation schemes suitable for transmission of voice and/or data over Doppler affected RiceChannel, which can be generated by this transmitter is given in Table 2.6. Details on modulation, coding and system issues can be found in Kucar (2006), Proakis (1983), Siegel (2001), Sklar (1988), and Van Trees (1968–1971).

Existing cellular radio systems such as AMPS, TACS, MCS, and NMT employ hybrid (analog and digital) schemes. For example, in access mode AMPS uses a digital codulation scheme (BCH coding and FSK modulation). While in information exchange mode, the frequency modulated analog voice is merged with discrete SAT and/or ST signals and occasionally blanked to send a digital message. These hybrid codulation schemes exhibit a constant envelope and as such allow the use of power efficient RF nonlinear amplifiers. On the receiver side, these schemes can be demodulated by an inexpensive, but efficient limiter/discriminator device. They require modest to high $C/N = 10$–20 dB, are very robust in adjacent (a spectrum is concentrated near the carrier) and cochannel interference (up to $C/I = 0$ dB, due

TABLE 2.6 Modulation Schemes: Glossary of Terms

Abbreviation	Description	Remarks/Use
ACSSB	Amplitude Companded Single SideBand	Satellite/terrestrial
AM	Amplitude Modulation	Broadcasting
APK	Amplitude Phase Keying modulation	
APM	Amplitude Phase Modulation	
ASK	Amplitude Shift Keying modulation	
BLQAM	Blackman (window) Quadrature Amplitude Modulation	
BPSK	Binary Phase Shift Keying	Spread spectrum systems, fiber
CPFSK	Continuous Phase Frequency Shift Keying	
CPM	Continuous Phase Modulation	
DEPSK	Differentially Encoded PSK (with carrier recovery)	
DPM	Digital Phase Modulation	
DPSK	Differential Phase Shift Keying (no carrier recovery)	
DSB-AM	Double SideBand AM	
DSB-SC-AM	Double SideBand Suppressed Carrier AM	Includes digital schemes
EQAM	Elliptic (window) Quadrature Amplitude Modulation	A subclass of DSB-SC-AM
FFSK	Fast Frequency Shift Keying ≡ MSK	NMT data and control
FM	Frequency Modulation	Broadcasting, AMPS voice
FQPSK	Feher QPSK	
FSK	Frequency Shift Keying	AMPS data and control
FSOQ	Frequency Shift Offset Quadrature modulation	
GMSK	Gaussian Minimum Shift Keying	GSM voice, data and control
GTFM	Generalized Tamed Frequency Modulation	
HMQAM	Hamming (window) Quadrature Amplitude Modulation	
IJF	Intersymbol Jitter Free ≡ SQORC	
LPAM	L-ary Pulse Amplitude Modulation	
LRC	LT symbols long Raised Cosine pulse shape	
LREC	LT symbols long Rectangularly EnCoded pulse shape	
LSRC	LT symbols long Spectrally Raised Cosine scheme	
MCPM	multi-*h* modulation index CPM	
MMSK	Modified Minimum Shift Keying ≡ FFSK	
MPSK	M-ary Phase Shift Keying	
MQAM	M-ary Quadrature Amplitude Modulation	A subclass of DSB-SC-AM
MQPR	M-ary Quadrature Partial Response	Radio-relay transmission
MQPRS	M-ary Quadrature Partial Response System ≡ MQPR	
MSK	Minimum Shift Keying ≡ FFSK	
NCFSK	NonCoherent FSK	
OFDM	Orthogonal Frequency Division Multiplexing modulation	DAB TV, ADSL
OQPSK	Offset (staggered) Quadrature Phase Shift Keying	
PM	Phase Modulation	Low capacity radio
PSK	Phase Shift Keying, 4PSK ≡ QPSK ≡ 4QAM	
QAM	Quadrature Amplitude Modulation	
QAPSK	Quadrature Amplitude Phase Shift Keying	
QORC	Quadrature Overlapped Raised Cosine	
QOSRC	Quadrature Overlapped Square Raised Cosine	
QPSK	Quadrature Phase Shift Keying ≡ 4 QAM	Radio satellite/terrestrial
QQPSK	Quadrature Quadrature Phase Shift Keying Q^2PSK	
SQAM	Staggered (offset) Quadrature Amplitude Modulation	
SQORC	Staggered Quadrature Overlapped Raised Cosine	
SQPSK	Staggered Quadrature Phase Shift Keying	
SSB	Single SideBand amplitude modulation	Low and high capacity radio
S3MQAM	Staggered class 3 Quadrature Amplitude Modulation	
TFM	Tamed Frequency Modulation	
TSI QPSK	Two-Symbol-Interval QPSK	
VSB	Vestigial SideBand	TV
WQAM	Weighted Quadrature Amplitude Modulation	Includes most digital schemes
XPSK	I/Q Crosscorrelated PSK	
$\pi/4$ DQPSK	$\pi/4$ shift DQPSK (with $\alpha = 0.35$ raised cosine filtering)	IS-54 TDMA voice and data
3MQAM	Class 3 Quadrature Amplitude Modulation	
4MQAM	Class 4 Quadrature Amplitude Modulation	
12PM3	12 state PM with 3 bit correlation	

Source: www.radio4u.com, 20061020.

to capture effect) cellular radio environment, and react quickly to the signal fade outages (no carrier, code, or frame synchronization). Frequency selective and Doppler affected mobile radio channels will cause modest to significant degradations known as the *random phase/frequency modulation.* By using modestly complex extended threshold devices C/N as low as 5 dB can provide satisfactory performance.

Tightly filtered codulation schemes, such as $\pi/4$ QPSK additionally filtered by a square root raised-cosine filter, exhibit a nonconstant envelope, which demands (quasi) linear, battery DC power inefficient amplifiers to be employed. On the receiver side, these schemes require complex demodulation receivers, a linear path for signal detection and a nonlinear one for reference detection—differential detection or carrier recovery. When such a transceiver operates in a selective fading multipath channel environment, additional countermeasures (inherently sluggish equalizers, etc.) are necessary to improve the performance—reduce the *bit error rate floor.* These codulation schemes require modest $C/N = 8$–16 dB and perform modestly in adjacent and/or cochannel (up to $C/I = 8$ dB) interference environment.

Codulation schemes employed in spread spectrum systems use low rate coding schemes and mildly filtered modulation schemes. When equipped with sophisticated amplitude gain control on the transmit and receive side, and robust rake receiver, these schemes can provide superior $C/N = 4$–10 dB and $C/I < 0$ dB performance. Unfortunately, a single transceiver has not been able to operate satisfactorily in a mobile channel environment. Consequently, a few additional signal codes and even an additional pilot tone have been employed to achieve required quality of the transmission. These pilot signals reduce significantly the spectrum efficiency in the forward direction and many times in the reverse direction. Furthermore, two combined QPSK like signals have up to (4×4) different baseband levels and may look like a 9QPR, or 16QAM signal, while three combined QPSK like signals may look like an MQPR signal, or like an MQAM signal. These combined signals, one information and two pilot signals, for example, at user's transmitter output, exhibit high peak factors and total power which is by 3–5 dB higher than C/N value necessary for a single information signal. In addition, inherently power inefficient linear RF power amplifiers are needed to be deployed; these three signal components of a CDMA scheme may have been optimized for minimal crosscorrelation and an easy of detection; as such, the same three signals may not necessarily have states in the QAM/QPR constellation that optimize peak-to-average ratio, and vice versa.

Enormous improvements in signal processing capabilities offered by ASIC and similar devices, while at the same time reducing power consumption, and availability of huge inexpensive memory devices, opened the doors for the development of complex coding schemes, interleavers, and algorithms; this includes sophisticated turbo coding and spatial coding for multiantenna systems. Powerful coding scheme, particularly useful in a broadcasting environment where processing delays are not important, provide low BER at C/N of only a few dB; however, these schemes may not be the best suitable for dynamic hostile channel environments. Transmitter and receiver diversity Multiple Input Multiple Output (MIMO) systems could offer significant enhancements in both capacity and quality of service, at the expense of system complexity. However, in a typical city cellular radio environment, macro diversity is by far superior to any micro diversity MIMO system.

2.11 Speech Coding

Human vocal tract and voice receptors, in conjunction with language redundancy (coding), are well suited for face to face conversation. As the channel changes (e.g., from telephone channel to mobile radio channel), different coding strategies are necessary to protect the loss of information.

In (analog) companded PM/FM mobile radio systems, speech is limited to 4 kHz, compressed in amplitude (2:1), preemphasized, and phase/frequency modulated. At a receiver, inverse operations are performed. Degradation caused by these conversions and channel impairments results in lower voice quality. Finally, the human ear and brain have to perform the estimation and decision processes on the received signal.

In digital schemes, sampling and digitizing of an analog speech (source) are performed first. Then, by using knowledge of properties of the human vocal tract and the language itself, a spectrally efficient source coding is performed. A high rate 64 kb/s, 56 kb/s, and AD-PCM 32 kb/s digitized voice complies with ITU-T recommendations for toll quality, but may be less practical for the mobile environment. One is primarily interested in 8–16 kb/s rate speech coders, which might offer satisfactory quality, spectral efficiency, robustness, and acceptable processing delays in a mobile radio environment. A glossary of the major speech coding schemes is provided in Table 2.7.

At this point, a partial comparison between analog and digital voice should be made. The quality of 64 kb/s digital voice, transmitted over a high-quality low noise telephone line, is essentially the same as the original analog voice (they receive nearly equal MOS). What does this *near equal* MOS mean in a radio environment? A mobile radio conversation consists of one (mobile to home) or a maximum of two (mobile to mobile) mobile radio paths, which dictate the quality of the overall connection. The results of a comparison between analog and digital voice schemes in different artificial mobile radio environments have been widely published. Generally, systems that employ digital voice and digital codulation schemes seem to perform well under modest conditions, while analog voice and analog codulation systems outperform their digital counterparts in fair and difficult (near threshold, in the presence of strong cochannel interference) conditions. Fortunately, present technology can offer a viable implementation of both analog and digital systems within the same mobile/portable radio telephone unit. This would give every individual a choice of either an analog or digital scheme, better service quality, and higher Customer Satisfaction. Trade-offs between the quality of digital speech, the complexity of speech and channel coding, as well as battery DC power consumption have to be assessed carefully, and compared with analog voice systems.

2.12 Macro and Micro Diversity

Macro diversity: Let us observe a typical evolution of a cellular system; in the beginning the base station may be located in the barycenter of the service area (center of the cell). The base station antenna is omnidirectional in azimuth, but with about 6–10 dBi gain in elevation, and serves most of the cell area, for example, >95%. Some parts within the cell may experience a lower quality of service because the direct path signal may be attenuated due to obstruction losses caused by buildings, hills, trees, and so on. The closest neighboring (the first tier) base stations serve corresponding neighboring areas cells by using different sets of frequencies, eventually causing adjacent channel interference. The second closest neighboring (the second tier) base stations might use the same frequencies (frequency reuse) causing cochannel interference. When the need for additional capacity arises and/or higher quality of service is required, the same nearly circular area may be divided into three 120° wide sectors, six 60° wide sectors, and so on, all served from the same base station location; now, the same base station is located at the edge of respective sectors. Since the new sectorial antennas provide 5 dB and 8 dB larger gains than the old omnidirectional antenna, respectively, these systems with new antennas with higher gains have longer spatial reach and may cover area belonging to neighboring cells of the old configuration; or, more probably, the output powers will be adjusted accordingly. For example, if the same real estate (base stations) is used in conjunction with 120° directional (in azimuth) antennas, the new designated 120° wide wedge area may be served by the previous base station and by two additional neighboring base stations now equipped with sectorial antennas with longer reach. Therefore, the same number of existing base stations equipped with new directional antennas and additional combining circuitry may be required to serve the same or different number of cells, yet in a different fashion. The mode of operation in which two or more base stations serve the same area is called the *macro diversity*. Statistically, three base stations are able to provide a better coverage of an area similar in size to the system with a centrally located base station. The directivity of a base station antenna (120° or even 60°) provides additional discrimination against signals from neighboring cells, therefore, reducing adjacent and cochannel interference, that is, improving reuse efficiency and capacity. Effective improvement depends on the terrain configuration, and the combining strategy and efficiency. However, it requires more complex antenna systems and combining devices.

Micro diversity employs two or more receptors at one site to received a desired signal. *Space diversity* systems employ two or more antennas spaced a distance apart from one another. A separation of only $\lambda/2 = 15$ cm at $f = 1$ GHz, which is suitable for implementation on the mobile side, can provide a notable

TABLE 2.7 Digital Vocoders: Glossary of Terms

Abbreviation	Description	Remarks/Use
ACELP	Adaptive CELP	
ADM	Adaptive Delta Modulation	
ADPCM	Adaptive Differential Pulse Code Modulation	Digital telephony 32 kb/s, DECT
ACIT	Adaptive Code sub-band excIted Transform	GTE
AMR	Adaptive MultiRate ACELP codec	dev.in 1999 for GSM, adopted for 3G
APC	Adaptive Predictive Coding	Inmarsat-B 16 kb/s
APC–AB	APC with Adaptive Bit allocation	
APC–HQ	APC with Hybrid Quantization	
APC–MQL	APC with Maximum Likelihood Quantization	
AQ	Adaptive Quantization	
ATC	Adaptive Transform Coding	
BAR	Backward Adaptive Reencoding	
CELP	Code Excited Linear Prediction	IS-95, U.S. Gov., half-rate GSM
CVSDM	Continuous Variable Slope Delta Modulation	
DAM	Diagnostic Acceptability Measure	
DM	Delta Modulation	A/D conversion
DPCM	Differential Pulse Code Modulation	
DRT	Diagnostic Rhyme Test	
DSI	Digital Speech Interpolation	TDMA FSS systems
DSP	Digital Signal Processing	
EVRC	Enhanced Variable Rate Speech codec	8 kb/s in CDMA Qualcom
HCDM	Hybrid Companding Delta Modulation	
IMBE	Inmarsat–M standard multiBand Excitation	6.4 kb/s, 4.15, 2.4 1.2 kb/s
JPAQ	Joint Pulse Amplitude Quantization	
LD–CELP	Low-Delay CELP	ITU-T G.728
LDM	Linear Delta Modulation	
LPC	Linear Predictive Coding	LPC-10 U.S. Gov. 2.4 kb/s, 4.8 kb/s
LTP	Long Term Prediction	
MBE	Multi Band Excitation	
MPLPC	Multi Pulse LPC	Skyphone aeronautical 9.6 kb/s
MSQ	Multipath Search Coding	
NIC	Nearly Instantaneous Companding	
PAME	Pitch Adaptive Mixed Excitation	
PCM	Pulse Code Modulation	Digital Voice, including 64 kb/s
PVXC	Pulse Vector Excitation Coding	
PWA	Predicted Wordlength Assignment	
QMF	Quadrature Mirror Filter	
RELP	Residual Excited Linear Prediction	GSM, Motorola STU III phones
RPE	Regular Pulse Excitation	RPE-LTP full-rate GSM 13 kb/s
SAVQ	Speaker Adaptive Vector Quantization	
SBC	Sub Band Coding	
SELP	Self Excitation Linear Prediction	
SIVP	Switched-adaptive Inter-frame Vector Prediction	
SMV	Selectable Mode Vocoder	CDMA Qualcom
STP	Short Term Prediction	
TASI	Time Assigned Speech Interpolation	TDMA FSS systems
TDHS	Time Domain Harmonic Scaling	
VAD	Voice Activity Detector	
VAPC	Vector Adaptive Predictive Coding	NASA MSAT-X
VCELP	Vector Code Excited Linear Prediction	
VEPC	Voice Excited Predictive Coding	
VQ	Vector Quantization	
VQL	Variable Quantum Level coding	
VSELP	Vector-Sum Excited Linear Prediction	IS-136, PDC, GSM half-rate 6.8 kb/s
VXC	Vector Excitation Coding	

Source: www.radio4u.com, 20061020.

improvement in some mobile radio channel environments. Micro space diversity is routinely used on cellular base sites. Macro diversity with the base stations located kilometers apart is also a form of space diversity.

Field-component diversity systems employ different types of antennas receiving either the electric or the magnetic component of an electromagnetic signal.

Frequency diversity systems employ two or more different carrier frequencies to transmit the same information. Statistically, the same information signal may or may not fade at the same time at different carrier frequencies. In a radionavigation satellite system, a frequency diversity system is also used for an estimation of characteristics of the propagation medium. FM, frequency hopping, and very wide band signaling can be viewed as frequency diversity techniques.

Time diversity systems are primarily used for the transmission of data. The same data are sent through the channel as many times as necessary, until the required quality of transmission is achieved—automatic repeat request (ARQ). *Would you please repeat your last sentence* is a form of time diversity used in a speech transmission.

The improvement of any diversity scheme is strongly dependent on the combining techniques employed, that is, the selective (switched) combining, the maximal ratio combining, the equal gain combining, the feedforward combining, the feedback (Granlund) combining, majority vote, hybrid, and so on, see Jakes (1974) and Simon (2000).

Continuous improvements in DSP and MMIC technologies and broader availability of ever improving CAD electromagnetics tools is making adaptive antennas solutions more viable than ever before. This is particularly through for systems above 1 GHz, where the same necessary base station antenna gain can be achieved with physically smaller antenna dimensions. An adaptive antenna could follow spatially shifting traffic patterns, adjust its gain and pattern, and consequently, improve the signal quality and capacity.

2.13 Broadcasting, Dispatch and Access

Signal paths have two distinct directions: the *forward link*—from the base station (via satellite) to all travelers within the footprint coverage area, and the *return link*—from a traveler (via satellite) to the base station. In the *forward link* direction a base station distributes information to travelers according to the previously established protocol, that is, no multiple access is involved; this way of operation is also called (*dispatch, distribution*) when information is delivered to a particular user, or *broadcasting* when information is delivered to a wide audience. In the *return link* direction many travelers make attempts to access one of the base stations; this way of operation is also called *access*. This occurs in so-called *control channels*, in a particular time slot, at particular frequency, or by using a particular code. If collisions occur, customers have to wait in a queue and try again until success is achieved. If successful, that is, no collision occurred, a particular customer will exchange (automatically) the necessary information for the call set-up. The network management center (NMC) will verify the customer's status, his credit rating, and so on; then, the NMC may assign a channel frequency, time slot, or code, on which the customer will be able to exchange information with his/her correspondent.

The optimization of the forward and reverse links may require different coding and modulation schemes and different bandwidths in each direction.

In forward link direction there are four basic distribution schemes: one which discriminates in frequency between different users is called *frequency division multiplex broadcasting* (FDMB), one which discriminates in time is called *time division multiplex broadcasting* (TDMB), one having different codes based on spread spectrum signaling is called *code division multiplex broadcasting* (CDMB), and a scheme having different frequency hopping codes is called *frequency hopping multiplex broadcasting* (FHMB). Hybrid schemes using a combination of basic schemes can also be used. All existing mobile radio communications systems employ FDM component; consequently, only FDMB schemes are pure, while the other three schemes are hybrid, that is, TDMB/FDM, CDMB/FDM, and FHMB/FDM; the hybrid solutions inherit complexities of both parents, that is, the need for an RF frequency synthesizer and a linear amplifier for *single channel per carrier* SCPC FDM solution, and the need for TDM and CDM overhead, respectively.

In reverse link direction there are four basic access schemes: one which discriminates in frequency between different users is called *frequency division multiple access* (FDMA), one which discriminates in time is called *time division multiple access* (TDMA), one having different codes based on spread spectrum signaling is called *code division multiple access* (CDMA), and a scheme having different frequency hopping codes is called *frequency hopping multiple access* (FHMA). Hybrid schemes using combination of basic schemes can also be developed.

A performance comparison of these schemes in a dynamic spatial environment is a very difficult task. The strengths of FDM schemes seem to be fully exploited in narrowband channel environments; to avoid the use of equalizers, channel bandwidths as narrow as possible should be employed; yet in such narrowband channels the quality of service is limited by the maximal expected Doppler frequency and practical stability of frequency sources; current limits are a few kHz.

The strengths of both TDM and CDM schemes seem to be fully exploited in wideband channel environments. TDM schemes need many slots, and bandwidth, to collect information on network behavior; once the equalization is necessary (at bandwidths >20 kHz), the data rate should be made as high as possible to increase frame efficiency, and freeze the frame to ease equalization; yet, high data rates require high-RF peak powers and a lot of signal processing power, which may be difficult to achieve in handheld units; current practical bandwidths are about 0.1–5 MHz. All existing schemes which employ TDM components are hybrid, that is, the TDMA/FDM schemes in which the full strength of the TDM scheme is not fully realized. CDM schemes need large spreading (processing) factors (and bandwidth) to realize spread spectrum potentials; however, high data rates require a lot of signal processing power, which may be difficult to achieve in handheld units; current practical bandwidths are up to about 20 MHz. A single code transceiver has not been able to operate satisfactorily in a hostile mobile channel environment; consequently, a few CDM elementary signals, that is, codes, information and pilot ones, may be necessary for successful transmission; this multi signal environment is powerwise similar to a MQAM/MQPR signaling scheme with a not necessarily optimal state constellation. Significant increase in the equipment complexity is accompanied with a significant increase in the average and peak transmitter power. In addition, a demanding RF synthesizer is needed to accommodate the CDMA/FDM mode of operation. Narrow frequency bands seem to favor FDM schemes, since both TDM and CDM schemes require more spectra to develop fully their potentials. However, once the adequate power spectrum is available, the later two schemes may be better suited for a complex (micro)cellular network environment. An inherently complex FHM system may be able to use different chunks of available spectra. Multiple access schemes are also delay and message sensitive. The length and type of message, and the type of service will influence the choice of multiple access, ARQ, frame and coding, among others.

2.14 System Capacity

The recent surge in the popularity of cellular radio, and mobile service in general, has resulted in an overall increase in traffic and a shortage of available system capacity in large metropolitan areas. Current cellular systems exhibit a wide range of traffic densities, from low in rural areas to overloading in downtown areas with large daily variations between peak hours and quiet night hours. It is a great system engineering challenge to design a system, which will make optimal use of the available frequency spectrum, offering a maximal traffic throughput (e.g., Erlangs/MHz/service area) at an acceptable service quality, constrained by the price and size of the mobile equipment. In a cellular environment, the overall system capacity in a given service area is a product of many (complexly interrelated) factors including the available frequency spectra, service quality, traffic statistics, type of traffic, type of protocol, shape and size of service area, selected antennas, diversity, frequency reuse capability, spectral efficiency of coding and modulation schemes, efficiency of multiple access, and so on.

In the 1970s, analog cellular systems employed omnidirectional antennas and simple diversity schemes offering modest capacity, that is, serving a modest number of customers. Analog cellular systems of the 1990s employ up to 60° sectorial antennas and improved diversity schemes; this combination results in a 3- to 5-fold increase in capacity. A further (twofold) increase in capacity from narrowband analog systems (25 kHz → 12.5 kHz) and nearly a threefold increase in capacity from the 5 kHz wide Narrowband AMPS; however, slight degradation in service quality might be expected. These improvements spurned the current growth in capacity, the overall success and prolonged life of analog cellular radio.

2.15 Conclusion

In this contribution, a broad repertoire of terrestrial and satellite systems and services for travelers is briefly described. This has been spiced with a brief historical overview of wire and wireless radio telegraph and mobile radio. The technical characteristics of the dispatch, cellular and cordless telephony systems are tabulated for ease of comparison. Issues such as operating environment, service quality, network complexity, cell size, channel coding and modulation (codulation), speech coding, macro and micro diversity, multiplex and multiple access, and the mobile radio communications system capacity are discussed. The use of frequency bands between 0.1 and 6.0 GHz by diverse radio communications and navigation systems is illustrated.

Presented data reveal significant differences between existing and planned terrestrial cellular mobile radio communications systems, and between terrestrial and satellite systems. These systems use different frequency bands, different bandwidths, different codulation schemes, different protocols, and so on, that is, they are not compatible.

What are the technical reasons for this incompatibility? In this contribution, performance dependence on multipath delay (related to the cell size and terrain configuration), Doppler frequency (related to the carrier frequency, data rate, and the speed of vehicles), and message length (may dictate the choice of multiple access) are briefly discussed. A system optimized to serve the travelers in the Great Plains may not perform very well in mountainous Switzerland; a system optimized for downtown cores may not be well suited to a rural environment; a system employing geostationary (above equator) satellites may not be able to serve travelers at high latitudes very well; a system appropriate for slow moving vehicles may fail to function properly in a high Doppler shift environment; a system optimized for voice transmission may not be very good for data transmission, and so on. A system designed to provide a broad range of services to everyone, everywhere, may not be as good as a system designed to provide a particular service in a particular local environment—as a decathlete world champion may not be as successful in competitions with specialists in particular disciplines.

There is plenty of opportunities where compatibility between systems, their integration, and frequency sharing may offer improvements in service quality, efficiency, cost and capacity, and therefore availability. Terrestrial systems offer a low start-up cost and a modest cost per user in densely populated areas. Satellite systems may offer a high quality of the service and may be the most viable solution to serve travelers in scarcely populated areas, on oceans, and in the air. Terrestrial systems are confined to two dimensions and radio propagation occurs in the near horizontal sectors. Barystationary satellite systems use the narrow sectors in the user's zenith nearly perpendicular to the Earth's surface having the potential for frequency reuse and an increase in the capacity in downtown areas during peak hours. A call set up in a forward direction (from the PSTN via base station to the traveler) may be a very cumbersome process in a terrestrial system when a traveler to whom a call is intended is roaming within an unknown cell. However, this may be realized easier in a global beam satellite system.

References

1. Akyldiz, I.F., Goodman, D.J., Kleinrock, L. (Guest Editors), 2001. Special Issue on Mobility and resource management in next generation wireless systems. *IEEE Journal on Selected Areas in Communications*, Vol. **19**, No.10, October 2001.
2. Ariyavisitakul, S.L., Falconer, D.D., Adachi, F., Sari, H. (Guest Editors), 1999. Special Issue on Broadband Wireless Techniques. *IEEE Journal on Selected Areas in Communications*, Vol. **17**, No.10, October 1999.

3. Cherubini, G., Cioffi, J.M., Duel-Hallen, A., Poor, H.V. (Guest Editors), 2001. Special Issue on Multiuser detection techniques with application to wired and wireless communications systems: Part I, *IEEE Journal on Selected Areas in Communications*, Vol. **19**, No.8, August 2001.
4. Cherubini, G., Cioffi, J.M., Duel-Hallen, A., Poor, H.V. (Guest Editors), 2001. Special Issue on Multiuser detection techniques with application to wired and wireless communications systems: Part II, *IEEE Journal on Selected Areas in Communications*, Vol. **20**, No.2, March 2002.
5. Chuang, J.C.-I., Anderson, J.B., Hattori, T., and Nettleton, R.W. (Guest Editors), 1993. Special Issue on Wireless Personal Communications: Part I, *IEEE Journal on Selected Areas in Communications*, Vol. **11**, No.6, August 1993.
6. Chuang, J.C.–I., Anderson, J.B., Hattori, T., and Nettleton, R.W. (Guest Editors), 1993. Special Issue on Wireless Personal Communications: Part II, *IEEE Journal on Selected Areas in Communications*, Vol. 11, No.7, September 1993.
7. Cimini, L.J., Tranter, W.H. (Guest Editors), 1999. Special Issue on Wireless Communications Series. *IEEE Journal on Selected Areas in Communications*, Vol. **17**, No.3, March 1999.
8. Cimini, L.J., Tranter, W.H. (Guest Editors), 1999. Special Issue on Wireless Communications Series. *IEEE Journal on Selected Areas in Communications*, Vol. **17**, No.7, July 1999.
9. Cimini, L.J., Tranter, W.H. (Guest Editors), 1999. Special Issue on Wireless Communications Series. *IEEE Journal on Selected Areas in Communications*, Vol. **17**, No.11, November 1999.
10. Cimini, L.J., Tranter, W.H. (Guest Editors), 1999. Special Issue on Wireless Communications Series. *IEEE Journal on Selected Areas in Communications*, Vol. **18**, No.3, March 2000.
11. Cox, D.C., Greenstein, L.J. (Guest Editors), 1995. Special Issue on Wireless Personal Communications. *IEEE Communications Magazine*, Vol. **33**, No.1, January 1995.
12. Cox, D.C., Hirade, K., Mahmoud, S.A. (Guest Editors), 1987. Special Issue on Portable and Mobile Communications. *IEEE Journal on Selected Areas in Communications*, Vol. **5**, No.4, June 1987.
13. Davis, J.H. (Guest Editor), Mikulski, J.J., Porter, P.T. (Associated Guest Editors), King, B.L. (Guest Editorial Assistant), 1984. Special Issue on Mobile Radio Communications. *IEEE Journal on Selected Areas in Communications*, Vol. **2**, No.4, July 1984.
14. Del Re, E., Devieux, C.L., Jr., Kato, S., Raghavan, S., Taylor, D., Ziemer, R. (Guest Editors), 1995. Special Issue on Mobile Satellite Communications for Seamless PCS. *IEEE Journal on Selected Areas in Communications*, Vol. **13**, No.2, February 1995.
15. Gabriel, William F. (Guest Editor), 1986. Special Issue on Adaptive Processing Antenna Systems. *IEEE Trans. on Antennas and Propagation*, Vol. **34**, No.3, March 1986.
16. Graglia, R.D., Luebbers, R.J., Wilton, D.R. (Guest Editors), 1997. Special Issue on Advanced Numerical Techniques in Electromagnetics. *IEEE Trans. on Antennas and Propagation*, Vol. **45**, No.3, March 1997.
17. Greenstein, L.J., Andersen, J.B., Bertoni, H.L., Kozono, S., Michelson, D.G., Tranter, W.H. (Guest Editors), 2002. Special Issue on Channel and propagation models for wireless system design I. *IEEE Journal on Selected Areas in Communications*, Vol. **20**, No.3, April 2002.
18. ION *Global Positioning System*, 1980–1993. Reprinted by The Institute of Navigation. Volume **I**. Washington, D.C., USA, 1980; Volume **II**. Alexandria, VA, USA, 1984; Volume **III**. Alexandria, VA, USA, 1986; Volume **IV**. Alexandria, VA, USA, 1993.
19. ITU, 1990. Recommendations of the CCIR, 1990 (also Resolutions and Opinions). *Mobile, Radiodetermination, Amateur and Related Satellite Services.* Volume **VIII**, **XVII**th Plenary Assembly, Düsseldorf, 1990. Reports of the CCIR, 1990 (also Decisions). *Land Mobile Service, Amateur Service, Amateur Satellite Service.* Annex 1 to Volume **VIII**, **XVII**th Plenary Assembly, Düsseldorf, 1990. Reports of the CCIR, 1990 (also Decisions). *Maritime Mobile Service.* Annex 2 to Volume **VIII**, **XVII**th Plenary Assembly, Düsseldorf, 1990. ITU-R Recommendations and Reports. CD-ROM 2005.
20. Jiangzhou, Wang, Adachi, F., Baier, P.W., Lehnert, J.S., Stark, W.E., Pursley, M.B. (Guest Editors), 2001. Special Issue on Wideband CDMA: Part I, *IEEE Journal on Selected Areas in Communications*, Vol. **18**, No.8, August 2000.
21. Jiangzhou, Wang, Adachi, F., Baier, P.W., Lehnert, J.S., Stark, W.E., Pursley, M.B. (Guest Editors), 2001. Special Issue on Wideband CDMA: Part II, *IEEE Journal on Selected Areas in Communications*, Vol. **19**, No.1, January 2001.
22. Kucar, A.D. (Guest Editor), 1991. Special Issue on Satellite and Terrestrial Systems and Services for Travelers. *IEEE Communications Magazine*, Vol. **29**, No.11, November 1991.

23. Kucar, A.D., Kato, S., Hirata, Y., Lundberg, O. (Guest Editors), 1992. Special Issue on Satellite Systems and Services for Travelers. *IEEE Journal on Selected Areas in Communications*, Vol. **10**, No.8, October 1992.
24. Kucar, A.D., Uddenfeldt, J. (Guest Editors), 1998. Special Issue on Mobile Radio Centennial. *Proceedings of the IEEE*, Vol. **86**, No.7, July 1998.
25. Mahmoud, S.A., Rappaport, S.S., Öhrvik, S.O. (Guest Editors), 1989. Special Issue on Portable and Mobile Communications. *IEEE Journal on Selected Areas in Communications*, Vol. **7**, No.1, January 1989.
26. Mailloux, R.J. (Guest Editor), 1999. Special Issue on Phased Arrays. *IEEE Trans. on Antennas and Propagation*, Vol. **47**, No.3, March 1999.
27. Mitola, J., III, Bose, V., Leiner, B.M., Turletti, T., Tennenhouse, D. (Guest Editors), 1999. Special Issue on Software Radios. *IEEE Journal on Selected Areas in Communications*, Vol. **17**, No.4, April 1999.
28. Oppermann, I., van Rooyen, P., Kohno, R. (Guest Editors), 1999–2000. Special Issue on Spread Spectrum for Global Communications: Part I, *IEEE Journal on Selected Areas in Communications*, Vol. **17**, No.12, December 1999.
29. Oppermann, I., van Rooyen, P., Kohno, R. (Guest Editors), 1999–2000. Special Issue on Spread Spectrum for Global Communications: Part II, *IEEE Journal on Selected Areas in Communications*, Vol. **18**, No.1, January 2000.
30. Radio Regulations, 1998 . . . 2004. International Telecommunication Union (ITU) Edition of 1998. Edition of 2001. Edition of 2004.
31. Rhee, S.B. (Editor), Lee, W.C.Y. (Guest Editor), 1991. Special Issue on Digital Cellular Technologies. *IEEE Trans. on Vehicular Technology*, Vol. **40**, No.2, May 1991.
32. Steele, R. (Guest Editor), 1992. Special Issue on PCS: The Second Generation. *IEEE Communications Magazine*, Vol. **30**, No.12, December 1992.
33. Shafi, M., Gesbert, D., Da-shan, Shiu, Smith, P.J., Tranter, W.H. (Guest Editors), 2003. Special Issue on MIMO systems and applications: Part I, *IEEE Journal on Selected Areas in Communications*, Vol. **21**, No.3, April 2003.
34. Shafi, M., Gesbert, D., Da-shan, Shiu, Smith, P.J., Tranter, W.H. (Guest Editors), 2003. Special Issue on MIMO systems and applications: Part II, *IEEE Journal on Selected Areas in Communications*, Vol. **21**, No.5, June 2003.
35. Siegel, P.H., Divsalar, D., Eleftheriou, E., Hagenauer, J., Rowitch, D., Tranter, W.H. (Guest Editors), 2001. Special Issue on The turbo principle: from theory to practice: Part I, *IEEE Journal on Selected Areas in Communications*, Vol. **19**, No.5, May 2001.
36. Siegel, P.H., Divsalar, D., Eleftheriou, E., Hagenauer, J., Rowitch, D., Tranter, W.H. (Guest Editors), 2001. Special Issue on The turbo principle: from theory to practice: Part II, *IEEE Journal on Selected Areas in Communications*, Vol. **19**, No.9, September 2001.
37. WARC, xxxx. *FINAL ACTS of the World Administrative Radio Conference*. WARC-79, Geneva, Switzerland, 1979, 0927–1205. WARC-92, Málaga–Torremolinos, Spain, 1992, 0131–0303.
38. WRC, xxxx. *FINAL ACTS of the World Radiocommunications Conference*. WRC-95, Geneva, Switzerland, 1995 1023–1117. WRC-97, Geneva, Switzerland, 1997, 1027–1121. WRC-00, Istanbul, Turkey, 2000, 0508–0602. WRC-03, Geneva, Switzerland, 2003, 0609–0704. WRC-07, Geneva, Switzerland, 2007, 1015–1109.
39. Ziolkowski, R.W., Engheta, N. (Guest Editors), 2003. Special Issue on Metamaterials. *IEEE Trans. on Antennas and Propagation*, Vol. **51**, No.10, October 2003.

For Further Information

The following trilogy, written by participants in AT&T Bell Laboratories projects on research and development in mobile radio, was, for the long time, during the golden age of engineering, the Holly Scripture of diverse cellular mobile radio topics:

Jakes, W.C. Jr. (Editor), 1974. *Microwave Mobile Communications*. John Wiley & Sons, Inc., New York, 1974.

AT&T Bell Labs Technical Personnel, 1979. Advanced Mobile Phone Service (AMPS). *Bell System Technical Journal*, Vol. **58**, No.1, January 1979.

Lee, W.C.Y., 1982. *Mobile Communications Engineering*. McGraw-Hill Book Company, New York, 1982.

An in-depth understanding of design, engineering and use of cellular mobile radio networks, including PCS and PCN, requires knowledge of diverse subjects such as three-dimensional cartography, electromagnetic propagation and scattering, computerized analysis and design of microwave circuits, fixed and adaptive antennas, analog and digital communications, project engineering, and so on. The following is a brief list of books relating to these topics.

Balanis Constantine, A., 1982, 1997. *Antenna Theory Analysis and Design*. Harper & Row, Publishers, New York, 1982. Second Edition, John Wiley & Sons, Inc., New York, 1997.

Bowman, J.J., Senior, T.B.A., Uslenghi, P.L.E., 1987. *Electromagnetic and Acoustic Scattering by Simple Shapes*. Revised Printing, Hemisphere Publishing Corporation, 1987.

Hansen, R.C., 1998. *Phased Array Antennas*. John Wiley & Sons, Inc., New York, 1998.

James, J.R., Hall, P.S. (Editors), 1989. *Handbook of Microstrip Antennas*. Volumes I and II. Peter Peregrinus Ltd., Great Britain, 1989.

Kucar, Andy D., 2001, 2006. *Satellite and Terrestrial Wireless Systems: Fixed and Mobile. PCS, PCN, LMDS, LAN, WAN, SDH, ATM, DSL, GPRS, 3G. Spread Spectrum. Radio versus Cable. A Practical Approach Toward the High Quality of Service*. Stridon Press, 2001. Second Edition, 2006.

Kucar, Andy D., 2006a. *Satellite Communications: An Introduction with Briefs on Frequency/Spectrum Management, Orbital Mechanics and Astronomy*. Stridon Press, 2006.

Kucar, Andy D., 2006b. *Time and Natural Calendars: From Kronos to GPS/Glonass > WAAS, MTSAT, EGNOS > GALILEO > GNSS*. Stridon Press, 2006.

Lee, Steve, 2002. *Spread Spectrum CDMA. IS–95 and IS–2000 for RF Communications*. McGraw-Hill Book Company, New York, 2002.

Lindsey, William, C., Marvin K. Simon, 1973. *Telecommunication Systems Engineering*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1973.

Lo, Y.T., Lee, S.W. (Editors), 1993. *The Antenna Handbook*. Volumes I–IV. Van Nostrand Reinhold, USA, 1993.

Mailloux, R.J., 1999. *Phased Array Antenna Handbook*. Artech House, Inc., Norwood, MA, 1994.

Proakis John, G., 1983. *Digital Communications*. McGraw-Hill Book Company, New York, 1983. Fourth Edition, McGraw-Hill Book Company, New York, 2000.

Silvester, P.P., Ferrari, R.L. 1996. *Finite Elements for Electrical Engineers*. Third Edition, Cambridge University Press, Great Britain, 1996.

Simon, Marvin K., Mohamed-Slim Alouini, 2000, 2005. *Digital Communication over Fading Channels: A Unified Approach to Performance Analysis*. John Wiley & Sons, Inc., New York, 2000. Second Edition, John Wiley & Sons, Inc., New York, 2005.

Sklar Bernard, 1988, 2001. *Digital Communications. Fundamentals and Applications*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1988. Second Edition, Prentice-Hall Inc., Englewood Cliffs, NJ, 2001.

Snyder John, P., 1987. *Map Projection—A Working Manual*. U.S. Geological Survey Professional Paper 1395, United States Government Printing Office, Washington, 1987. (Second Printing 1989).

Spilker, J.J., Jr., 1977. *Digital Communications by Satellite*. Prentice-Hall Inc., Englewood Cliffs, NJ, 1977.

Stutzman Warren, L., Gary, A. Thiele, 1981, 1998. *Antenna Theory and Design*. John Wiley & Sons, Inc., New York, 1981. Second Edition, John Wiley & Sons, Inc., New York, 1998.

Van Trees Harry, L., 1968, 1971. *Detection, Estimation, and Modulation Theory*. Part I, 1968, Part II, 1971, Part III, John Wiley & Sons, Inc., New York, 1971.

Walker, J., 1999. *Advances in Mobile Information Systems*. Artech House, Inc., Norwood, MA, 1999.

Web Sites

http://europa.eu.int/comm/dgs/energy_transport/galileo/index_en.htm
http://gps.faa.gov/
http://wi-fi.org/OpenSection/index.asp
http://www.3gpp.org/
http://www.alcatel.com/
http://www.ansi.org/
http://www.arib.or.jp/english/
http://www.att.com/
http://www.bluetooth.com/
http://www.consumer.philips.com/
http://www.cospas-sarsat.org/
http://www.ericsson.com/
http://www.etsi.org/
http://www.fcc.gov/
http://www.globalstar.com/
http://www.glonass-center.ru/frame_e.html
http://www.gsmworld.com/
http://www.ic.gc.ca/cmb/welcomeic.nsf/icPages/Menu-e
http://www.iec.ch/
http://www.ieee.org/
http://www.inmarsat.com/
http://www.irda.org/
http://www.iridium.com/
http://www.itu.int/home/index.html
http://www.lenovo.com.cn/
http://www.motorola.com/
http://www.nist.gov
http://www.nokia.com/
http://www.nortel.com/
http://www.nttdocomo.co.jp/english/p_s/service/foma/index.html
http://www.oas.org/EN/PROG/CITEL/citel.htm
http://www.optus.com.au/
http://www.orbcomm.com/
http://www.panasonic.com/
http://www.qualcomm.com/
http://www.samsung.com/
http://www.siemens.de/
http://www.sirius.com/
http://www.thuraya.com/
http://www.tiaonline.org/standards/overview.cfm
http://www.tdscdma-forum.org/EN/index.asp
http://www.tta.or.kr/English/new/main/index.htm
http://www.wimaxforum.org/home/
http://www.xmradio.com/
http://www.zigbee.org/en/index.asp

3

Cellular Mobile Telephony

Paul G. Flikkema
Northern Arizona University

3.1 Introduction

The goal of modern cellular mobile telephone systems is to provide wireless communication services to people as efficiently as possible. In the past, this definition would have been restricted to mobile telephone services. However, in new telephone service markets, the cost of wireless infrastructure is less than wired infrastructure. Thus wireless mobile telephony technology is being adapted to provide in-home telephone service, the so-called wireless local loop (WLL). Indeed, with approximately 2 billion subscribers worldwide in 2006, it appears that wireless telephony will become dominant over traditional wired access worldwide. More importantly, nonvoice services are proliferating rapidly, and mobile telephone technology is now being integrated into a wide array of personal, portable devices.

The objective of this chapter is to familiarize the RF/microwave engineer with the concepts and terminology of cellular mobile telephony ("cellular"), or mobile wireless networks. A capsule history and a summary form the two bookends of the chapter. In between, we start with the cellular concept and the basics of mobile wireless networks. Then we take a look at some of the standardization issues for cellular systems. Following that, we cover the focus of the standards battles: channel access methods. We then take a look at some of the basic aspects of cellular important to RF/microwave engineers: first, modulation, diversity, spread spectrum, and orthogonal frequency-division multiplexing (OFDM); then coding, interleaving, and time diversity; and, finally, nonlinear channels. Before wrapping up, we take a glimpse at a topic of growing importance, namely, antenna array technology.

3.2 A Brief History

Mobile telephone service was inaugurated in the United States in 1946 with six radio channels available per city. This evolved into the manual Mobile Telephone System (MTS) used in the 1950s and 1960s. The year 1964 brought the Improved MTS (IMTS) systems with eight channels per city with—finally—no telephone operator required. Later, the capacity was more than doubled to 18. Most importantly, the IMTS introduced narrowband frequency modulation (NBFM) technology. The first cellular service was introduced in 1983, called AMPS (Advanced Mobile Phone Service). Cities were covered by cells averaging about 1 km in radius, each served by a base station. This system used the 900 MHz frequency band still in use for mobile telephony. The cellular architecture allowed frequency reuse, dramatically increasing capacity to a maximum of 832 channels per cell.

The age of digital, or second generation, cellular did not arrive until 1995 with the introduction of the IS-54 TDMA service and the competing IS-95 CDMA service. In 1996–1997 the U.S. Federal Communications Commission auctioned licenses for mobile telephony in most U.S. markets in the so-called PCS (Personal Communication System) bands at 1.9 GHz. These systems use a variety of standards, including TDMA (time division multiple access), CDMA (code division multiple access), and the GSM TDMA standard that originated in Europe. Outside the United States, a similar evolution has occurred, with GSM deployed in Europe and the PDC (personal digital cellular) system in Japan. In other countries there has been a pitched competition between all systems. While not succeeding in the United States so-called low-tier systems have been popular in Europe and Japan. These systems are less robust to channel variations and are therefore targeted to indoor residential and commercial use in Europe, and became popular for data services and children in Japan. The European system is called DECT (Digital European Cordless Telephony) and the Japanese system is called PHS (Personal Handyphone System).

Third-generation (or 3G) mobile telephone service was first demonstrated in 2000, but market acceptance has been slow and subscribers will be converting to the new services in 2005–2010. To maximize return on investment in infrastructure, service providers have been filling the gap between 2G and 3G with so-called 2.5G services, including packet data overlay services and improved modulation. The 3G services will be offered in the context of a long-lived standardization effort recently renamed IMT-2000 (International Mobile Telecommunications-2000) under the auspices of the Radio Communications Standardization Sector of the International Telecommunications Union (ITU-R; see http://www.itu.int). Key goals of IMT-2000 are as follows [1]:

1. Use of a common frequency band over the globe
2. Worldwide roaming capability
3. Higher transmission rates than second-generation systems to handle new data-over-cellular applications

Another goal is to provide the capability to offer asymmetric rates, so that the subscriber can download data much faster than he/she can send it.

While also aiming for worldwide access and roaming, the main technical thrust of 3G systems is to provide high-speed wireless data services—up to 5 Mbps for downlink data (incoming to the user). However, from the service providers' perspective, the goal is to provide a vast array of high-speed data services to personal mobile devices, including IP-based services (e.g., voice over IP), network gaming, video telephony, push-to-talk, web browsing, and other news and information services.

3.3 The Cellular Concept

At first glance, a logical method to provide radio-based communication service to a metropolitan area is a single centrally located antenna (base station). However, radio frequency spectrum is a limited commodity, and regulatory agencies, in order to meet the needs of a vast number of applications, have further limited the amount of RF spectrum for mobile telephony. The limited amount of allocated spectrum forced designers to adopt the cellular approach: using multiple base stations to cover a geographic area; each base station covers a roughly circular area called a cell. Figure 3.1 shows how a large region can be split into seven smaller cells (approximated by hexagons). This allows different base stations to use the same frequencies for communication links as long as they are separated by a sufficient distance. This is known as frequency reuse, and allows thousands of mobile telephone users in a metropolitan area to share far fewer channels.

There is a second important aspect to the cellular concept. With each base station covering a smaller area, the mobile phones need less transmit power to reach any base station (and thus be connected with the telephone network). This a major advantage, since, with battery size and weight a major impediment to miniaturization, the importance of reducing power consumption of mobile phones is difficult to overestimate.

If two mobile units are connected to their respective base stations at the same frequency (or more generally, channel), interference between them, called *cochannel interference*, can result. Thus there is a trade-off between frequency reuse and signal quality, and a great deal of effort has resulted in frequency assignment techniques that balance this trade-off. They are based on the idea of clustering: taking the available set of channels, allocating them in chunks to each cell, and arranging the cells into geographically local clusters. Figure 3.2 shows how clusters of seven cells (each allocated one of seven mutually exclusive channel subsets) are used to cover a large region; note that the arrangement of clusters maximizes the reuse distance—the distance between any two cells using the same frequency subset. Increasing the reuse distance has the effect of reducing cochannel interference.

Although it might seem attractive to make the cells smaller and smaller, there are diminishing returns. First, smaller cell sizes increase the need for management of mobile users as they move about. In other words, smaller cell sizes require more handoffs, where the network must transfer users between base stations. Another constraint is antenna location, which is often limited by available space and esthetics. Fortunately, both problems can be overcome by technology. Greater handoff rates can be handled by increases in processing speed, and creative antenna placement techniques (such as on lamp posts or sides of buildings) are allowing higher base station densities.

Another issue is evolution: How can a cellular system grow with demand? Two methods have been successful. The first is cell splitting: by dividing a cell into several cells (and adjusting the reuse pattern), a cellular service provider can increase its capacity in high-demand areas. The second is sectoring: instead of a single omnidirectional antenna covering a cell, a typical approach is to sectorize the cell into N_S regions,

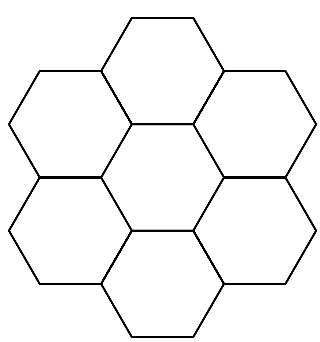


FIGURE 3.1 A region divided into cells. While normally the base stations are placed at the center of the cells, it is also possible to use edge-excited cells where base stations are placed at vertices.

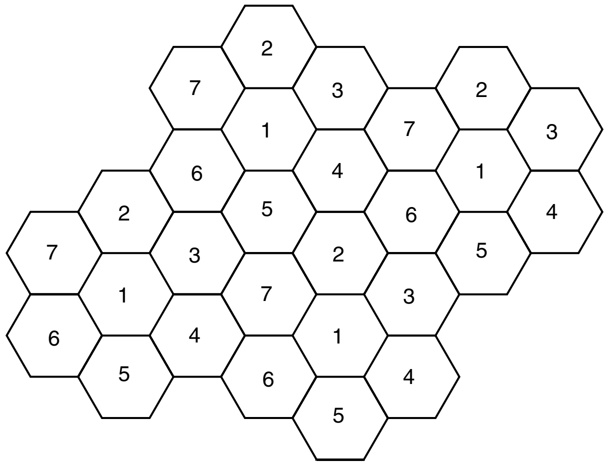


FIGURE 3.2 Cell planning with cluster size of 7. The number indexes the subset of channels used in the cell. Other cluster sizes, such as 4, 7, or 12 can be used.

each served by an antenna that covers an angular span of $2\pi/N_S$ ($N_S = 3$ is typical). Note that both approaches increase handoff rates and thus require concurrent upgrading of network management. Later we will describe smart antennas, the logical extension to sectorization.

3.4 Networks for Mobile Telephony

A communication network that carries only voice—even a digital one—is relatively simple. Other than the usual digital communication system functions, such as channel coding, modulation, and synchronization, all that is required is call setup and takedown. However, current and future digital mobile telephony networks are expected to carry digital data traffic as well.

Data traffic is by nature computer-to-computer, and requires that the network has an infrastructure that supports everything from the application (such as web browsing) to the actual transfer of bits. The data is normally organized into chunks called packets (instead of streams as in voice), and requires a much higher level of reliability than digitized voice signals. These two properties imply that the network must also label the packets, and manage the detection and retransmission of packets that are received in error. It is important to note that packet retransmission, while required for data to guarantee fidelity, is not possible for voice since it would introduce delays that would be intolerable in a human conversation.

Other functions that a modern digital network must perform include encryption and decryption (for data security) and source coding and decoding. The latter functions minimize the amount of the channel resource (in essence, bandwidth) needed for transferring the information. For voice networks this involves the design of voice codecs (coder/decoders) that not only digitize voice signals but also strip out the redundant information in them. In addition to all the functions that wired networks provide, wireless networks with mobile users must also provide mobility management functions that keep track of calls as subscribers move from cell to cell.

The various network functions are organized into *layers* to rationalize the network design and to ease internetworking, or the transfer of data between networks [2]. RF/microwave engineering is part of the lowest, or *physical layer* that is responsible for carrying the data over the wireless medium.

3.5 Standards and Standardization Efforts

The cellular industry is, if anything, dense with lingo and acronyms. Here, we try to make sense of at least some of the important and hopefully longer-lived terminology.

Worldwide, most of the cellular services are offered in two frequency bands: 900 and 1900 MHz. In each of the two bands, the exact spectrum allocated to terrestrial mobile services varies from country to country. In the United States, cellular services are in the 800–900 MHz band, while similar services are in the 880–980 MHz band in Europe under the name GSM900 (GSM900 combines in one name a radio communication standard—GSM, or Global System for Mobile Communications—and the frequency band in which it is used. We will describe the radio communication, or *air interface*, standards later). In the mid-1990s, the United States allocated spectrum for PCS (Personal Communication Services) from 1850 to 2000 MHz; while many thought PCS would be different from cellular, they have converged and are interchangeable from the customer's perspective. Similarly, in Europe GSM1800 describes cellular services offered using the 1700–1880 MHz band.

The 1992 World Administrative Radio Conference (WARC'92) allocated spectrum for third-generation mobile radio in the 1885–1980 and 2110–2160 MHz bands. The ITU-R's IMT-2000 standardization initiative adopted these bands for terrestrial mobile services. Note that the IMT-2000 effort is an umbrella that includes both terrestrial and satellite-based services—the latter for areas for where terrestrial services are unavailable.

There are two major industry-driven standardization efforts as of 2006: 3GPP (Third-Generation Partnership Project, 3gpp.org) is dedicated to the GSM standard and its evolution into W-CDMA (wideband CDMA), while 3GPP2 (3gpp2.org) works on the competing CDMA2000 standard. Both bodies are working to develop technological roadmaps to high-speed wireless data services. Note that this information is subject to change in future WARCs and in evolving standards and their organizations.

The cellular *air interface* standards are designed to allow different manufacturers to develop both base station and subscriber (mobile user handset) equipment. They are generally different for the downlink (base station to handset) and uplink (handset to base station). This reflects the asymmetry of resources available: the handsets are clearly constrained in terms of power consumption and antenna size, so that the downlink standards imply sophisticated transmitter design, while the uplink standards emphasize transmitter simplicity and advanced receive-side algorithms. The air interface standards address channel access protocols as well as traditional communication link design parameters such as modulation and coding. Looking ahead, the major research challenges are what designers can achieve in terms of both higher bandwidths and higher spectral efficiency. These concepts are taken up in the following sections.

3.6 Channel Access

In a cellular system, a fixed amount of RF spectrum must somehow be shared among thousands of simultaneous phone conversations or data links. *Channel access* is about (1) dividing the allocated RF spectrum into pieces and (2) allocating the pieces to conversations/links in an efficient way.

The easiest channel access method to envision is FDMA (Frequency Division Multiple Access), where each link is allocated a subband (i.e., a specific carrier frequency; see Figure 3.3). This is exactly the access method used by first generation (analog) cellular systems. The second generation of cellular brought two newer channel access methods that were enabled by progress in digital processing technology. One is TDMA, wherein time is divided into *frames*, and links are given short *time slots* in each frame (Figure 3.4). FDMA and TDMA can be seen as time/frequency duals, in that FDMA subdivides the band into narrow subbands in the frequency domain, while TDMA subdivides time into slots, during which a link (within a cell) uses the entire allocated bandwidth.

The second generation of cellular also brought CDMA. In CDMA, all active links simultaneously use the entire allocated spectrum, but sophisticated codes are used that allow the signals to be separated in the receiver.* We will describe CDMA in more depth later.

* It is fashionable to depict CDMA graphically using a "code dimension" that is orthogonal to the time–frequency plane, but this is an unfortunate misrepresentation. Similar to any signals, CDMA signals exist (in fact, overlap) in the time–frequency plane, but have correlation-based properties that allow them to be distinguished.

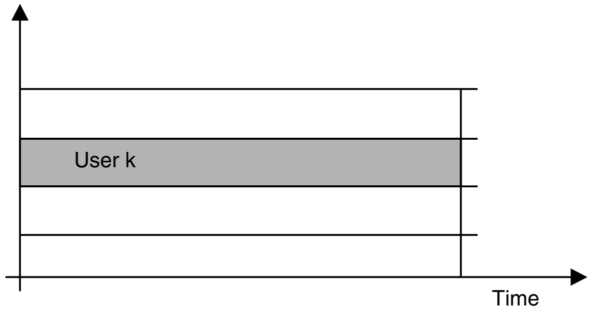


FIGURE 3.3 FDMA depicted on the time–frequency plane, with users assigned carrier frequencies, or channels. Not shown are guard bands between the channels to prevent interference between users' signals.

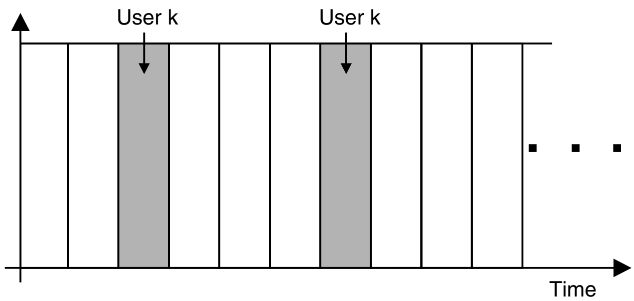


FIGURE 3.4 Depiction of TDMA on the time–frequency plane. Users are assigned time slots within a frame. Guard times (not shown) are needed between slots to compensate for timing inaccuracies.

It should be noted that both TDMA- and CDMA-based cellular systems also implicitly employ FDMA, though this is rarely mentioned. The reason is that the cellular bands are divided into smaller bands (a form of FDMA), and both TDMA and CDMA are used within these subbands.

In the United States, the TDMA and CDMA standards are referred to by different acronyms. The TDMA standard originally was called IS-54, but with enhancements became IS-136. The CDMA standard was called IS-95, and has been rechristened as cdmaOne by its originator, Qualcomm. These standards were created under the auspices of the Telecommunications Industry Association (TIA) and the Electronic Industries Alliance (EIA).

In Europe, the second generation brought digital technology in the form of the GSM standard, which used TDMA. (The GSM standard originally referred to Group Special Mobile, but was updated to capture its move to worldwide markets.) Japan also chose TDMA in its first digital offering, called PDC.

The three multiple access approaches use different signal properties (frequency, time, or code) to allow distinguishing of multiple signals. How do they compare? In the main, as we move from FDMA to TDMA to CDMA (in order of their technological development), complexity is transferred from the RF section to the digital section of the transmitters and receivers. The evolution of multiple access techniques has tracked the rapid evolution of digital processing technology as it has become cheaper and faster. For example, while FDMA requires a tunable RF section, both TDMA and CDMA need only a fixed-frequency front end. CDMA relieves one requirement of TDMA—strict synchronization among the various transmitters—but introduces a stronger requirement for synchronization of the receiver to the high-bandwidth received signal. In addition, the properties of the CDMA signal provide a natural means to exploit the multipath nature of the digital signal for improved performance. However, these advantages come at the cost of massive increases in the capability of digital hardware. Luckily, Moore's Law (i.e., that processing power roughly doubles every 18 months at similar cost) still remains in effect as of the turn of the century, and the amount of processing power that will be used in the digital phones in the twenty-first century will be unimaginable to the architects of the analog systems developed in the 1970s.

3.7 Modulation

The general purpose of modulation is to transform an information-bearing message signal into a related signal that is suitable for efficient transmission over a communication channel. In *analog modulation*, this is a relatively simple process: the information-bearing analog (or continuous-time) signal is used to alter a parameter (normally, the amplitude, frequency, or phase) of a sinusoidal signal (or carrier, the signal carrying the information). For example, in the NBFM modulation used in the 1980s AMPS system, the voice signal alters the frequency content of the modulated signal in a straightforward manner.

The purpose of *digital modulation* is to convert an information-bearing discrete-time symbol sequence into a continuous-time waveform. Digital modulation is easier to analyze than analog modulation, but more difficult to describe and implement.

3.7.1 Modulation in Digital Communication

Before digital modulation of the data in the transmitter, there are several processing steps that must be applied to the original message signal to obtain the discrete-time symbol sequence. A continuous-time message signal, such as the voice signal in telephony, is converted to digital form by sampling, quantization, and source coding. *Sampling* converts the original continuous-time waveform into discrete-time format, and *quantization* approximates each sample of the discrete-time signal using one of a fixed number of levels. Then *source coding* jointly performs two functions: it strips redundancy out of the signal and converts it to a discrete-time sequence of symbols.

What if the original signal is already in discrete-time (sampled format), such as a computer file? In this case, no sampling or quantization is needed, but source coding is still used to remove redundancy.

Between source coding and modulation is a step critical to the efficiency of digital communications: channel coding. This is discussed later; it suffices for now to know that it converts the discrete-time sequence of symbols from the source coder into another (better) discrete-time symbol sequence for input to the modulator. Following modulation, the signal is upconverted, filtered (if required), and amplified in RF electronics before being sent to the antenna. All the steps described are shown in block diagram form in Figure 3.5. In the receiver, the signal from the antenna, following filtering (again, if required), is amplified and downconverted prior to demodulation, channel decoding, and source decoding (see Figure 3.5).

What is the nature of the digital modulation process? The discrete-time symbol sequence from the channel coder is really a string of symbols (letters) from an *finite* alphabet. For example, in *binary* digital modulation, the input symbols are 0s and 1s. The modulator output converts those symbols into one of a *finite* set of waveforms that can be optimized for the channel.

While it is the finite set of waveforms that distinguishes digital modulation from analog modulation, the difference is only one manifestation of the entire paradigm of *digital communication*. In a good digital communication design, the source coder, channel coder, and modulator all work together to maximize the efficient use of the communication channel; even two of the three are not enough for good performance.

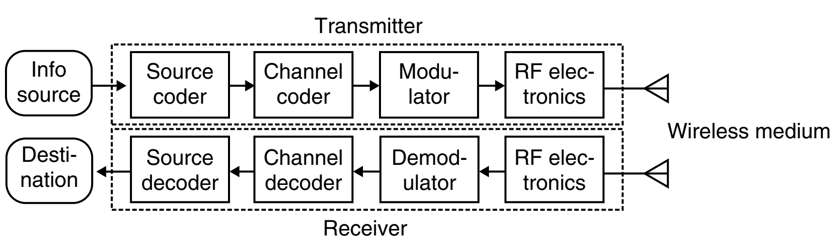


FIGURE 3.5 Communication system block diagram for wireless communication. In the wireless medium, multipath propagation and interference can be introduced. For system modeling purposes, the two blocks of RF electronics are combined with the wireless medium to form the wireless channel—a channel that distorts the signal, and adds noise and interference. The other blocks are designed to maximize the system performance for the channel.

3.7.2 Selection of Digital Modulation Formats

There are several (often conflicting) criteria for the selection of a modulation scheme. They are

- BER (bit error rate) performance
  - in wireless, particularly in cellular mobile channels, the scheme must operate under conditions of severe fading.
  - cellular architectures imply cochannel interference.
  - Typically, 10^{-2} or better is required for voice telephony and 10^{-5} or better is required for data.
- Spectral (or bandwidth) efficiency (measured in bits/s/Hz)
- Power efficiency (especially for handheld/mobile terminals)
- Implementation complexity and cost

In the U.S. cellular market, complexity is of special importance: with the number of standards growing, many handsets are now dual- and triple-mode; for example, a phone might have both GSM and 3G capability. While some hardware can be shared, multimode handsets clearly place additional constraints on the allowable complexity for each mode.

3.7.3 Classification of Digital Modulation Schemes

Broadly, modulation techniques can be classified into two categories.

Linear methods include schemes that use combinations of amplitude and phase modulation of a pulse stream. They have higher spectral efficiencies than constant-envelope methods (see following), but must use more-expensive (or less efficient) linear amplifiers to maintain performance and to limit out-of-band emissions.

Examples of linear modulation schemes include PSK (phase-shift keying) and QAM (quadrature amplitude modulation). QAM can be viewed as a generalization of PSK in that both the amplitude and the phase of the modulated waveform are altered in response to the input symbols.

Constant-envelope methods are more complicated to describe, but usually are sophisticated methods based on frequency modulation. Their key characteristic is a constant envelope (resulting in a constant instantaneous signal power) regardless of the source symbol stream. They allow use of less expensive amplification and/or higher amplification efficiencies (e.g., running amplifiers in the nonlinear region), at the expense of out-of-band emissions. Historically, they are limited to spectral efficiencies of about 1 bits/s/Hz.

Examples of constant envelope methods include FSK (frequency-shift keying) and more sophisticated methods such as MSK and GMSK (these will be described shortly). These methods can be thought of as digital (finite alphabet) FM in that the spectrum of the output signal is varied according to the input symbol stream.

The spectral occupancy of a modulated signal (per channel) is roughly

$$S_{\mathrm{O}} = B + 2\Delta f,$$

where B is the bandwidth occupied by signal power spectrum and Δf is the maximum one-way carrier frequency drift.* We can express the bandwidth as

$$B = \frac{R_{\mathrm{d}}}{\epsilon},$$

where R_{d} is the channel data rate (in bits/s) and ϵ is the spectral efficiency (in bits/s/Hz). Combining,

$$S_{\mathrm{O}} = \frac{R_{\mathrm{d}}}{\epsilon} + 2\Delta f.$$

*This drift can be caused by oscillator instability or Doppler due to channel time variations.

Thus, to minimize spectral occupancy (thus maximizing capacity in number of users), we can employ one of the following steps:

1. Reduce R_d by lowering the source coding rate (implying more complexity or lower fidelity)
2. Improve spectral efficiency of modulation (implying higher complexity)
3. Improve transmitter/receiver oscillators (at greater cost)

3.7.4 Modulation, Up/Downconversion, and Demodulation

To transmit a string of binary information symbols (or bits—zeros and ones), $\{b_0, b_1, b_2, \ldots\}$, we can represent a 1 by a positive-valued pulse of amplitude one, and a 0 by a negative pulse of the sample amplitude. This mapping from the bit value at time n, b_n, to amplitude a_n can be accomplished using

$$a_n = 2b_n - 1.$$

To complete the definition, we define a pulse of unit amplitude with start time of zero and stop time of T as $p_T(t)$. Then the modulated signal can be efficiently written as

$$u(t) = \sum_n a_n p_T(t - nT).$$

This signal is at baseband—centered at zero frequency—and is therefore unsuitable for wireless communication media. However, this signal can be upconverted to a desired RF by mixing with a sinusoid to get the passband signal

$$x(t) = u(t)\cos(2\pi f_c t) = \cos(2\pi f_c t)\sum_n a_n p_T(t - nT),$$

where f_c is the carrier frequency.

Multiplying a sinusoid by ± 1 is identical to changing its phase between 0 and π radians, so we have

$$x(t) = \cos\left(2\pi f_c t + \sum_n d_n p_T(t - nT)\right),$$

where we assign $d_n = 0$ when $a_n = -1$ and $d_n = \pi$ when $a_n = 1$. This equation shows that we are simply shifting the phase of the carrier between two different values: this is BPSK (binary phase-shift keying).

Why not use more than two phase values? In fact, four are ordinarily used for better efficiency: pairs of bits are mapped to four different phase values 0, $\pm\pi/2$, and π. For example, the CDMA standards employ this scheme, known as quaternary PSK (QPSK).

In general, the baseband signal will be complex-valued, which leads to the general form of upconversion from baseband to passband:

$$x(t) = \sqrt{2}\,\Re\{u(t)\,e^{j2\pi f_c t}\},$$

where the $\sqrt{2}$ factor is simply to maintain a consistency in measurement of signal power between passband and baseband. The motivation of using the baseband representation of a signal is twofold: first, it retains the amplitude and phase of the passband signal, and is thus independent of any particular carrier frequency; second, it provides the basis for modern baseband receiver implementations that use high-speed digital signal processing. The baseband representation is also known as the *complex envelope* representation.

BPSK and QPSK are linear modulation methods; in contrast, FSK is a constant-envelope modulation scheme. For binary FSK (BFSK), there are two possible signal pulses, given at baseband by

$$u_0(t) = A\,\mathrm{e}^{-j\pi\Delta f t}\,p_T(t), \quad u_1(t) = A\,\mathrm{e}^{j\pi\Delta f t}\,p_T(t),$$

where A is the amplitude. Notice that we have two (complex) tones separated by Δf. MSK (minimum-shift keying) and GMSK (Gaussian pre-filtered MSK) are special forms of FSK that provide greater spectral efficiency at the cost of higher implementation efficiency. The GSM standard uses GMSK, while its enhanced 2.5G version, EDGE (Enhanced Data Rates for Global Evolution), uses GMSK for voice and low-rate data and an eight-phase PSK modulation scheme (achieving 3 bits per symbol) for higher rates.

At the receiver, the RF signal is amplified and downconverted with appropriate filtering to remove interference and noise. The downconverted signal is then passed to the demodulator, whose function is to detect (guess in an optimum way) what symbol stream was transmitted. Following demodulation (also referred to as detection), the symbol stream is sent to subsequent processing steps (channel decoding and source decoding) before delivery to the destination.

At this point it is typical to consider the BER and spectral efficiencies of various digital modulation formats, modulator and demodulator designs, and the performance of different detection strategies for mobile cellular channels. This is beyond the scope of this chapter, and we direct the reader to a good book on digital communications (e.g., [3–7]) for more information. A more recent book [8] provides a good introduction to newer topics, including turbo codes, adaptive algorithms, and networks.

3.8 Diversity, Spread Spectrum, and CDMA

A mobile wireless channel causes the transmitted signal to arrive at the receiver via a number of paths due to reflections from objects in the environment. If the channel is linear (including transmit and receive amplifiers), a simple modeling approach for this multipath channel is to assume that it is specular, that is, each path results in a specific amplitude, time delay, and phase change. If the channel is also at least approximately time-invariant, its impulse response under these conditions can be expressed as*

$$h(t) = \sum_{\lambda=0}^{\Lambda} \alpha_\lambda\,\mathrm{e}^{j\theta_\lambda}\,\delta(t-\tau_\lambda),$$

where α_λ, τ_λ, and θ_λ are, respectively, the amplitude, time delay, and phase for the λ-th path.

Let the transmitted signal be

$$s(t) = \sum_n a_n f_n(t),$$

a sequence of pulses $f_n(t)$ each modulated by a transmitted symbol a_n at a symbol rate of $1/T$. When transmitted via a specular multipath channel with Λ paths, the received signal—found by the convolution of the transmitted signal and the channel impulse response—is

$$y(t) = \sum_{\lambda=0}^{\Lambda} \alpha_\lambda\,\mathrm{e}^{j\theta_\lambda}\,s(t-\tau_\lambda),$$

For simplicity, consider sending only three symbols a_{-1}, a_0, a_1. Then the received signal becomes

$$y(t) = \sum_{n=-1}^{1} a_n \sum_{\lambda=0}^{\Lambda} \alpha_\lambda\,\mathrm{e}^{j\theta_\lambda} f_n(t-\tau_\lambda).$$

* Here $\delta(t)$ denotes the Dirac delta function.

Two effects may result—fading and intersymbol interference. *Fading* occurs when superimposed replicas of the same symbol pulse nullify each other due to phase differences. *Intersymbol interference* (ISI) is caused by the convolutive mixing of the adjacent symbols in the channel. Fading and ISI may occur individually or together depending on the channel parameters and the symbol rate T^{-1} of the transmitted signal.

Let us consider in more detail the case where the channel *delay spread* is a significant fraction of T, that is τ_Λ is close to, but smaller than T. In this case we can have both fading and ISI, which, if left untreated, can severely compromise the reliability of the communication link. Direct-sequence spread-spectrum (DS/SS) signaling is a technique that mitigates these problems by using clever designs for the signaling pulse $f_n(t)$. These pulse designs are wide bandwidth (hence "spread spectrum"), and the extra bandwidth is used to endow them with properties that allow the receiver to separate the symbol replicas.

Suppose we have two-path channel, and consider the received signal for symbol a_0. Then the DS/SS receiver separates the two replicas

$$\alpha_0\, e^{j\theta_0} a_0 f_0(t-\tau_0), \quad \alpha_1\, e^{j\theta_1} a_0 f_0(t-\tau_1).$$

Then each replica is adjusted in phase by multiplying it by $\alpha_\lambda\, e^{-j\theta_\lambda}$, $\lambda = 0, 1$ yielding (since $zz^* = |z|^2$)

$$\alpha_0 a_0 f(t-\tau_0), \quad \alpha_1 a_0 f(t-\tau_1).$$

Now all that remains is to delay the first replica by $\tau_1 - \tau_0$ so they line up in time, and sum them, which gives

$$(\alpha_0 + \alpha_1) a_0 f(t-\tau_1).$$

Thus DS/SS can turn the multipath channel to advantage—instead of interfering with each other, the two replicas are now added constructively. This *multipath combining* exploits the received signal's inherent *multipath diversity*, and is the basic idea behind the technology of RAKE reception* used in the CDMA digital cellular telephony standards.

It is important to note that this is the key idea behind all strategies for multipath fading channels: we somehow exploit the redundancy, or *diversity* of the channel (recall the multiple paths). In this case, we used the properties of DS/SS signaling to effectively split the problematic two-path channel into two benign one-path channels. Multipath diversity can also be viewed in the frequency domain, and is in effect a form of *frequency diversity*. As we will see later, frequency diversity can be used in conjunction with other forms of diversity afforded by wireless channel, including time diversity and antenna diversity.

CDMA takes the spread spectrum idea and extends it to the separation of signals from multiple *transmitters*. To see this, suppose M transmitters are sending signals simultaneously, and assume for simplicity that we have a single-path channel. Let the complex (magnitude/phase) gain for channel m be denoted by $\beta^{(m)}$. Finally, the transmitters use different spread-spectrum pulses, denoted by $f^{(m)}(t)$. If we just consider the zeroth transmitted symbols from each transmitter, we have the received signal

$$y(t) = \sum_{m=1}^{M} \beta^{(m)} a_0^{(m)} f^{(m)}(t - t_m),$$

where the time offset t_m indicates that the pulses do not necessarily arrive at the same time.

The above equation represents a complicated mixture of the signals from multiple transmitters. If narrowband pulses are used, they would be extremely difficult—probably impossible—to separate. However, if the pulses are spread-spectrum, then the receiver can use algorithms to separate them from each other,

*The RAKE nomenclature can be traced to the block diagram representation of such a receiver—it is reminiscent of a garden rake.

and successfully demodulate the transmitted symbols. Of course, these ideas can be extended to many transmitters sending long strings of symbols over multipath channels.

Why is it called CDMA? It turns out that the special properties of the signal pulses $f^{(m)}(t)$ for each user (transmitter) m derive from high-speed *codes* consisting of periodic sequences of chips $c_k^{(m)}$ that modulate chip waveforms $\varphi(t)$. One way to envision it is to think of $\varphi(t)$ as a rectangular pulse of duration $T_c = T/N$. The pulse waveform for user m can then be written

$$f^{(m)}(t) = \sum_{k=0}^{N-1} c_k^{(m)} \varphi(t - kT_c).$$

The fact that we can separate the signals means that we are performing code-division multiple access—dividing up the channel resource by using codes. Recall that in FDMA this is done by allocating frequency bands, and in TDMA, time slots. The pulse waveforms in CDMA are designed so that many users' signals occupy the entire bandwidth simultaneously, yet can still be separated in the receiver. The signal-separating capability of CDMA is extremely important, and can extend beyond separating desired signals within a cell. For example, the IS-95 CDMA standard uses spread-spectrum pulse designs that enable the receiver to reject a substantial amount of cochannel interference (interference due to signals in other cells). This gives the IS-95 system (as well as its proposed 3G descendents) its well-known property of universal frequency reuse.

The advantages of DS/SS signals derive from what are called their *deterministic correlation* properties. For an arbitrary periodic sequence $\{c_k^{(m)}\}$, the deterministic *autocorrelation* is defined as

$$\phi^{(m)}(i) = \frac{1}{N} \sum_{k=0}^{N-1} c_k^{(m)} c_{k+i}^{(m)},$$

where i denotes the relative shift between two replicas of the sequence. If $\{c_k^{(m)}\}$ is a direct-sequence spreading code,

$$\phi^{(m)}(i) \approx \begin{cases} 1, & i = 0 \\ 0, & 0 < |i| < N. \end{cases}$$

This "thumbtack" autocorrelation implies that relative shifts of the sequence can be separated from each other. Noting that each chip is a fraction of a symbol duration, we see that multipath replicas of a symbol pulse can be separated even if their arrival times at the receiver differ by less than a symbol duration.

CDMA signal sets also exhibit special deterministic *cross-correlation* properties. Two spreading codes $\{c_k^{(l)}\}$, $\{c_k^{(m)}\}$ of a CDMA signal set have the cross-correlation property

$$\phi^{(l,m)}(i) = \frac{1}{N} \sum_{k=0}^{N-1} c_k^{(l)} c_{k+i}^{(m)} \approx \begin{cases} 1, & l = m, i = 0, \\ 0, & l = m, 0 < |i| < N, \\ 0, & l \neq m. \end{cases} \tag{3.1}$$

Thus we have a set of sequences with zero cross-correlations and "thumbtack" autocorrelations. (Note that this includes the earlier autocorrelation as a special case.) The basic idea of demodulation for CDMA is as follows: if the signal from user m is desired, the incoming received signal—a mixture of multiple transmitted signals—is correlated against $\{c_k^{(m)}\}$. Thus multiple replicas of a symbol from user m can be separated, delayed, and then combined, while all other users' signals (i.e., where $l \neq m$) are suppressed by the correlation.

Details of these properties, their consequences in demodulation, and descriptions of specific code designs can be found in References 5, 6, 9, and 10.

3.9 Orthogonal Frequency-Division Multiplexing

OFDM is a technique that, like CDMA, captures aspects of modulation and multiple access. Recall that in CDMA, each information symbol is "carried" by multiple chips in time. In OFDM, a high-speed information stream is split into multiple lower-speed streams. Each of these streams is then placed into a certain frequency subband such that the collection of subbands contains the entire signal. In this way, the information stream is now modulated onto multiple carriers. However, from an RF perspective, these can be considered "pseudo" carriers, since they are actually created using digital signal processing: an inverse fast fourier transform (FFT) converts the collection of subband signals into a single time-domain signal that modulates a single carrier. In the receiver, a forward FFT after downconversion recovers the subbands, each of which can be demodulated.

Because it is not required to load every subband with information, OFDM can easily support variable bit rates. For this reason, OFDM is seen as a very likely candidate for 4G (or "beyond 3G") systems because in combination with TDMA it allows the time-frequency plane to be chopped up into different size chunks, thus providing a convenient mechanism to dynamically support very diverse application bit rates [11]. Thus, a communication link can easily be upgraded from a low bit rate (e.g., to support a traditional voice call) to a high bit-rate link that could support on-demand video. OFDM has traditionally been used to combat distortion large multipath-induced delay spreads by subdividing a short symbol into a number of longer symbols. This advantage has already been exploited in the OFDM-based 802.11a/n high speed wireless LAN standards. Of course, hybrid systems using OFDM with other schemes, such as CDMA and multiple antennas, are possible and have significant potential [12,13]; see also Reference 14.

3.10 Channel Coding, Interleaving, and Time Diversity

As we have mentioned, channel coding is a transmitter function that is performed after source coding, but before modulation. The basic idea of channel coding is to introduce structured redundancy into the signal that will allow the receiver to easily detect or correct errors introduced in the transmission of the signal.

Channel coding is fundamental to the success of modern wireless communication. It can be considered the cornerstone of digital communication, since, without coding, it would not be possible to approach the fundamental limits established by Shannon's information theory [15,16].

The easiest type of channel codes to understand are *block codes*: a sequence of input symbols of length k is transformed into a code sequence (codeword) of length $n > k$. Codes are often identified by their rate R, where $R = k/n \leq 1$. Generally, codes with a lower rate are more powerful. Almost all block codes are *linear*, meaning that the sum of two codewords is another codeword. By enforcing this linear structure, coding theorists have found it easier to find codes that not only have good performance, but also have reasonably simple decoding algorithms as well.

Error-correcting codes have enough power so that errors can actually be corrected in the receiver. Systems that use these codes are called *forward error-control* (FEC) systems. *Error-detecting codes* are simpler, but less effective: they can tell *whether* an error has occurred, but not where the error is located in the received sequence; hence, it cannot be corrected.

Error-detecting codes can be useful when it is possible for the receiver to request retransmission of a corrupted codeword. Systems that employ this type of feedback are called ARQ (Automatic Repeat-reQuest systems).

In wireless systems, *convolutional codes* have been popular. Instead of blocking the input stream into length-k sequences and encoding each one independently, convolutional coders are finite-state sequential machines. Therefore they have memory, so that a particular output symbol is determined by a contiguous sequence of input symbols. For example, a rate-1/2 convolutional coder outputs two code symbols for each information symbol that arrives at its input. Normally, these codes are also linear.

As we have seen, the fading in cellular systems is due to multipath. Of course, as the mobile unit and other objects in the environment move, the physical structure of the channel changes with time, causing the fading of the channel to vary with time. However, this fading process tends to be slow relative to the symbol rate, so a long string of coded symbols can be subjected to a deep channel fade. In other words, the fading from one symbol to the next will be highly correlated. Thus, the fades can cause a large string of demodulation (detection) errors or an *error burst*. Thus, fading channels are often described from the point of view of coding as *burst-error channels*.

The most well-known block and convolutional codes are best suited to random errors, that is, errors that occur in an uncorrelated fashion and thus tend to occur as isolated single errors. Although there have been a number of codes designed to correct burst errors, the theory of random error-correcting codes is so well developed that designers have often chosen to use these codes in concert with a method to "randomize" error bursts.

This randomization method, called *interleaving*, rearranges, or scrambles, the coded symbols in order to minimize this correlation so that errors are isolated and distributed across a number of codewords. Thus, a modest random-error correcting code can be combined with interleaving to efficiently handle burst errors. In the transmitter, the interleaver is inserted between the channel coder and the modulator to shuffle the symbols of the code words. Then, in the receiver, the deinterleaver is placed between the demodulator and the decoder to reassemble the codewords for decoding.

We note that a well-designed coding/interleaving system does more than redistribute errors for easy correction: it also exploits *time diversity*. In our discussion of spread-spectrum and CDMA, we saw how the DS/SS signal exploits the frequency diversity of the wireless channel via its multipath redundancy. Here, the redundancy added by channel coding/interleaving is designed so that, in addition to the usual performance increase due to just the code—the *coding gain*—there is also a benefit to distributing the redundancy in such a way that exploits the time variation of the channel, yielding a *time diversity gain*.

In this era of digital data over wireless, high link reliability is required. This is in spite of the fact that most wireless links have a raw bit error rate (BER) on the order of 1 in 1000. Clearly, we would like to see an error rate of 1 in 10^{12} or better. How is this astounding improvement achieved? The following two-level approach has proved successful. The first level employs FEC to correct a large percentage of the errors. This code is used in tandem with a powerful error-detecting algorithm to find the rare errors that the FEC cannot find and correct. This combined FEC/ARQ approach limits the amount of feedback to an acceptable level while still maintaining the necessary reliability.

For 3G mobile systems, new channel codes discovered in the 1990s are being implemented. They are remarkable in that in a very real sense, their discovery closes a chapter on communication theory opened by Shannon by nearly achieving channel capacity for noise-only point-to-point links. These codes are usually classified into two types, known as *turbo codes* and *low-density parity check* (LDPC) codes. Although they are quite different in their origins and implementations, they are based on the same concepts. First, they use large near-random codewords. Second, decoding methods are based on iterative statistical inference of the transmitted bits, building confidence in a potential decision at each step; the final decision is thus made with knowledge about the reliability of the decision.

3.11 Nonlinear Channels

Amplifiers are more power-efficient if they are driven closer to saturation than if they are kept within their linear regions. Unfortunately, nonlinearities that occur as saturation is approached lead to *spectral spreading* of the signal. This can be illustrated by observing that an instantaneous (or memoryless) nonlinearity can be approximated by a polynomial. For example, a quadratic term effectively squares the signal; for a sinusoidal input this leads to double-frequency terms.

A more sophisticated perspective comes from noting that the nonlinear amplification can distort the symbol pulse shape, expanding the spectrum of the pulse. Nonlinearities of this type are said to cause

AM/AM distortion. Amplifiers can also exhibit AM/PM conversion, where the output phase of a sinusoid is shifted by different amounts depending on its input power—a serious problem for PSK-type modulations.

A great deal of effort has gone into finding transmitter designs that allow more efficient amplifier operation. For example, constant-envelope modulation schemes are insensitive to nonlinearities, and signaling schemes that reduce the peak-to-average power ratio (PAPR) of the signal allow higher levels. Finally, methods to linearize amplifiers at higher efficiencies are receiving considerable attention.

Modeling and simulating nonlinear effects on system performance is a nontrivial task. AM/AM and AM/PM distortions are functions of frequency, so if wideband amplifier characterization is required, a family of curves is necessary. Even then the actual wideband response is only approximated, since these systems are limited in bandwidth and thus have memory. More accurate results in this case can be obtained using Volterra series expansions, or numerical solutions to nonlinear differential equations. Sophisticated approaches are becoming increasingly important in cellular as supported data rates move higher and higher. More information can be found in References 3 and 17, and the references therein.

3.12 Antenna Arrays

We have seen earlier how sectorized antennas are used to increase system performance. They are one of the most economical forms of multielement antenna systems, and can be used to reduce interference or to increase user capacity. A second use of multielement systems is to exploit the *spatial diversity* of the wireless channel. Spatial diversity approaches assume that the receive antenna elements are immersed in a signal field whose strength varies strongly with position due to a superposition of multipath signals arriving via various directions. The resulting element signal strengths are assumed to be at least somewhat statistically uncorrelated. This spatial uncorrelatedness is analogous to the uncorrelatedness over time or frequency that is exploited in mobile channels. In essence, the use of antenna arrays allows the addition *spatial diversity* to a menu that already includes time and frequency diversity.

Now rarely used, the simplest approaches employ multiple (nominally omnidirectional in azimuth) antenna elements at the receiver, and choose the one with the highest signal-to-noise ratio. More sophisticated schemes combine—rather than select just one of—the element signals to further improve the signal-to-noise ratio at the cost of higher receiver complexity. These approaches date from the 1950s, and do not take into account other interfering mobile units. These latter schemes are often grouped under the category of *antenna diversity* approaches.

More recently, techniques for systems that integrate multiple antennas with signal processing and coding are already being introduced. The first is *spatial multiplexing*, wherein independent bit streams are transmitted from each antenna, and multiantenna signal processing techniques are used in the receiver to demultiplex the signals. The second area of research combines multiple antennas with error-control coding mechanisms under the name of *space-time coding* [18], and more recently space-frequency coding using OFDM [13]. The main contribution of these efforts has been the recognition that multiple-element transmit and receive antennas can, under certain conditions, dramatically increase the link capacity via massive increases in spectral efficiency.

Another approach, beamforming or phased-array antennas, is also positioned to play a role in future systems under the new moniker *smart antennas*. Space-time coding and beamforming methods can be seen as two approaches to exploit the properties of systems with multiple transmit and receive antennae, often called multiple-input/multiple-output (MIMO) systems. However, in contrast to space-time coding approaches, strong inter-element correlation based on the DOA of plane waves is assumed. Beamforming employs an array of antenna elements connected to an amplitude- and phase-shifting (or time delay) network to adaptively tune (steer electronically) the antenna pattern based on the angles of arrival of the signals.

There is a great deal of activity to determine optimum methods that integrate multiple antennas with signal processing techniques for modulation, multiple access, spatial multiplexing, coding, and beamforming. The ultimate goal of these approaches can be stated as follows: to track individual mobile units with optimized antenna patterns that maximize performance (by maximizing the ratio of the signal to the sum of interference and noise), minimize power consumption at the mobile unit, and optimize the capacity of the cellular system. One can conjecture that the ultimate solution to this problem will be a class of techniques that involve joint design of adaptive channel coding, modulation, and antenna array processing [19,20].

3.13 Summary

Almost all wireless networks are distinguished by the characteristic of a shared channel resource, and this is in fact the key difference between wireless and wired networks. Another important difference between wired and wireless channels is the presence of multipath in the latter, which makes diversity possible. What is it that distinguishes cellular from other wireless services and systems? First, it historically has been designed for mobile telephone users, and was optimized for carrying human voice. This has led to the following key traits of cellular:

- Efficient use of spectrum via the cellular concept.
- System designs, including channel access mechanisms, that efficiently handle large numbers of uniform—that is, voice—links.
- Difficult channels: user mobility causes fast variations in channel signal characteristics compared with other wireless applications such as wireless local area networks.

The second trait has reached its end, as now wireless voice is just one of many applications that will be designed into phones and numerous other products. Thus mobile telephony will become just one of many radio access networks (RANs) that will complement and compete with other RANs.

We close by mentioning an apparent trend that may have a significant impact on wireless/RF engineering. The first is the integration of multiple wireless technologies in a single portable device [21]. For example, current smartphones include Wi-Fi/802.11 for high-speed wireless local area networking (WLAN), Bluetooth for wireless personal area networking (WPAN), and GPS for location services, all in addition to cellular service. Other services, such as WiMAX (high-speed subscriber wireless) and digital audio broadcast, seem inevitable as well. These developments may drive new design approaches that exploit both commonality and reconfigurability [22] at device and subsystem levels in the continuing quest for smaller and lighter products.

References

1. M. Zeng, A. Annamalai, and V. K. Bhargava. Recent advances in cellular wireless communications. *IEEE Communications Magazine*, 37(9):128–138, September 1999.
2. J. Walrand and Pravin Varaiya. *High-Performance Communication Networks*. Morgan Kaufman, 1996.
3. J. B. Anderson. *Digital Transmission Engineering*. IEEE Press, 1999.
4. E. A. Lee and D. G. Messerschmitt. *Digital Communication*. Kluwer Academic, 2nd edition, 1994.
5. J. G. Proakis. *Digital Communications*. McGraw-Hill, 3rd edition, 1995.
6. S. Haykin. *Communication Systems*. Wiley, 1994.
7. J. G. Proakis and M. Salehi. *Communication Systems Engineering*. Prentice-Hall, 1994.
8. A. Goldsmith. *Wireless Communications*. Cambridge University Press, 2005.
9. P. Flikkema. Introduction to spread spectrum for wireless communication: a signal processing perspective. *IEEE Signal Processing Magazine*, 14(3):26–36, May 1997.
10. A. J. Viterbi. CDMA: *Principles of Spread Spectrum Communication*. Addison-Wesley, 1995.
11. H. Ekström et al. Technical solutions for the 3G long-term evolution. *IEEE Communications Magazine*, 44(3):38–45, 2006.
12. V. Chakravarthy et al. TDCS, OFDM, and MC-CDMA: a brief tutorial. *IEEE Communications Magazine*, 43(9):S11–S16, 2005.
13. M. Juntti et al. MIMO MC-CDMA communications for future cellular systems. *IEEE Communications Magazine*, 43(2):118–124, 2005.

14. A. Jamalipour, T. Wada, and T. Yamazato. A tutorial on multiple access technologies for beyond 3G mobile networks. *IEEE Communications Magazine*, 43(2):110–117, 2005.
15. C. E. Shannon. Communication in the presence of noise. *Proceedings of the IRE*, 37(1):10–21, January 1949.
16. A. D. Wyner and S. Shamai (Shitz). Introduction to "Communication in the presence of noise" by C. E. Shannon. *Proceedings of the IEEE*, 86(2):442–446, February 1998. Reprinted in the *Proceedings of the IEEE*, 86(2):447–457, February 1998.
17. M. C. Jeruchim, P. Balaban, and K. S. Shanmugan. *Simulation of Communication Systems*. Plenum, 1992.
18. J. Mietzner and P. Hoeher. Boosting the performance of wireless communication systems: theory and practice of multiple-antenna techniques. *IEEE Communications Magazine*, 42(10):40–47, 2004.
19. A. Alexiou and M. Haardt. Smart antenna technologies for future wireless systems: trends and challenges. *IEEE Communications Magazine*, 42(9):90–97, 2004.
20. D. Tse and P. Viswanath. *Fundamentals of Wireless Communication*. Cambridge University Press, 2005.
21. S. Frattasi et al. Defining 4G technology from the user's perspective. *IEEE Network*, 20:35–41, 2006.
22. K. Muhammad, R. Staszewski, and D. Leipold. Digital RF processing: toward low-cost reconfigurable radios. *IEEE Communications Magazine*, 43(8):105–113, 2005.

4

WiMAX RF System and Circuit Challenges

Balvinder Bisla
Intel Corporation

Broadband Wireless Access has occupied a niche in the market for about a decade, but with the signing of the 802.16d standard it could finally explode into the mass market. With the emergence of this standard an ecosystem is developing that will allow multiple vendors to produce components that adhere to a standard specification and hence allow large-scale deployment. One of the major challenges of the 802.16d standard is the plethora of options that exist; Worldwide Interoperability Microwave Access (WiMAX) will address this issue by limiting options and hence ensure interoperability. The result will allow manufacturers of radio frequency (RF) components and test equipment to have their products used for mass deployment.

In this chapter, we focus on the various RF challenges that exist on a RF system level and show how such challenges can translate into circuit designs. The RF is made more complicated by the fact that WiMAX indeed addresses wireless markets across the world both in licensed and unlicensed bands. Thus, solutions have to be flexible enough to allow for the many RF bands and different regulations around the globe. Several major RF architectures are discussed and the implications for WiMAX specifications are explored; in particular, both intermediate frequency (IF)- and I/Q-based structures are investigated.

Part of our discussion will provide insight into the cost and performance trade-offs between time division duplex (TDD) and frequency division duplex (FDD) systems both in licensed and unlicensed bands. It is generally accepted that TDD systems offer cost advantages over their FDD counterparts; however, most licensed bands intended for data applications operate with FDD systems in mind. Some of the RF subsystem blocks that have stringent WiMAX specifications are also elaborated upon: these include synthesizers, power amplifiers (PA), and filters. These fundamental subsystem blocks are where most of the transceiver costs reside; the same blocks are also responsible for most of the RF performance.

The industry is moving toward using Orthogonal Frequency Division Multiplexing Access (OFDMA) and either spatial diversity or beamforming techniques to enhance link margins. We touch on the RF challenges associated with these techniques. Finally, we view some of the important WiMAX specifications for RF and the implications for the design of RF circuits, which include SNDR, channel bandwidths, RF bands, noise figures, output power levels, and gain setting. Some important differences between WiMAX and 802.11 RF specifications are also highlighted.

4.1 Introduction

As the RF challenges mount, so do the costs of the Radio. For WiMAX to be successful the cost versus performance equation has to be balanced carefully. Two extreme examples of this cost and performance equation are a single in single out (SISO) system from Hybrid Networks (now defunct) requiring line of sight (LOS) radios. LOS radios result in truck rollouts utilizing experienced technicians to set the equipment up. However, the cost of the radio is low due to its simplicity. In general, the SISO radio requires expensive installation and reliability is poor; link margins are typically 145 dB. On the other hand, Iospan Wireless (now defunct) demonstrated a multiple in multiple out (MIMO) radio with a 3×2 system; that is, three Receive and two Transmit chains. It was able to support link margins of 165 dB that could penetrate inside homes in multipath environments. With this ability, the issue of costly truck rollouts is eliminated; however, the cost of the multiple radio chains becomes a deterrent. Still, as radio frequency integrated circuit (RFIC) integration improves, costs will head down. WiMAX, through the use of integration and advanced techniques to increase link margins, should be able to achieve reliable wireless systems at a reasonable cost.

4.2 RF Architectures

This section describes the plethora of trade-offs and challenges for RF architectures for WiMAX-related radios. We discuss FDD and its cousin Half FDD (HFDD) as well as TDD. IF, direct conversion or zero intermediate frequency (ZIF), as well as variants of these are presented. The interface between the Baseband (BB) chip and the radio must be carefully designed, so these challenges are exposed. Methods to improve link margins, namely MIMO, and beamforming can be used in WiMAX. In addition, OFDMA, which allows for subchannelization, improves capacity efficiency. We discuss the RF challenges inherent in the use of these methods.

4.3 TDD/FDD and HFDD Architectures

4.3.1 TDD

Figure 4.1 shows a TDD radio. The darkened blocks are the most costly in the radio.

TDD systems utilize one frequency band for both Transmit and Receive. This concept requires only one local oscillator (LO) for the radio. In addition, only one RF filter is necessary and this filter is shared between the Transmitter (TX) and the Receiver (RX). The synthesizer and RF filters are major cost drivers in radios. Having one synthesizer saves on die area; a large part of the radio die size can be taken up by the LO, in particular the inductor, which is part of the resonant structure.

The RF filter in a TDD system is not required to attenuate its TX Noise as severely as in FDD systems. The TDD mode prevents the TX Noise from self jamming the RX since only one is on at any time. As well as relief of the RF filter specifications, having just one RF filter saves cost and space. It should be noted that to ensure Transmitting radios do not interfere with nearby Receiving radios, the specification for TX

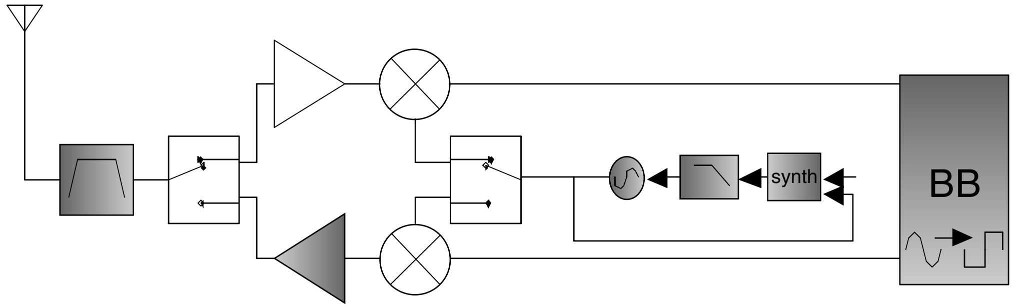


FIGURE 4.1 TDD radio.

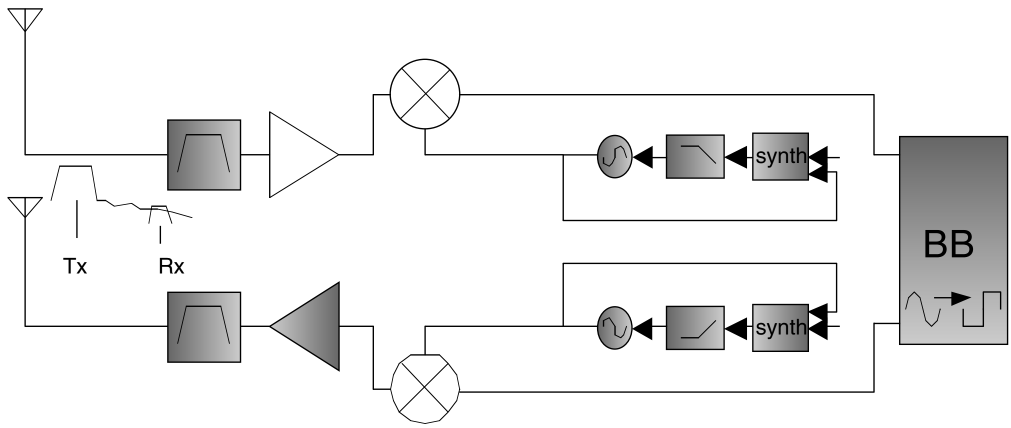


FIGURE 4.2 FDD radio.

Noise cannot be eased with abandon. The transmission noise from Radio 1 will interfere with the received signal of Radio 2. Thus, although self jamming specifications are made easier, collocation specifications must be carefully considered. There is a notable savings in power from the TDD architecture, a direct result of turning the RX off while in TX mode and vice versa.

Several disadvantages exist, however. There is a reduction of data throughput since there is no transmission of data while in RX mode unlike FDD systems. The medium access control (MAC) level software tends to have a more complicated scheduler than an FDD system since it must deal with synchronizing many users' time slots in both TX and RX mode. It must be noted that while the RF filtering specifications are relaxed, this tends to imply that subscriber stations will have to be spaced further apart from each other to avoid interference. In essence, the system must handle fewer users in a given area than in FDD systems.

TDD systems are most prominent in unlicensed bands; in these bands, the regulations for output noise are more relaxed than in licensed bands. Thus inexpensive RF filters can be specified. As the unlicensed bands are free of cost, there is a competition to drive for the lowest cost architecture, TDD.

4.3.2 FDD

Figure 4.2 shows an FDD radio.

A high-performance RF front-end is required in FDD systems. Collocation issues from a TX Noise perspective are solved since the worst-case scenario of self jamming is not possible. FDD systems do not have to switch the RX or TX; this alleviates settling time specifications, which results in a simpler radio design. The MAC software is simpler because it does not have to deal with the time synchronization issues as in TDD systems.

The radio must be capable of data transmission while in Receive mode without incurring any degradation in bit error rate (BER). To ease the burden on the filter there is a gap between the TX frequency band and the RX band; however, carriers wish to minimize this space. Typically this is a separation of 50–100 MHz.

We try to specify the TX Noise to be 10 dB below the RX input noise floor, in which case the TX Noise will only degrade the RX by 0.5 dB. Unfortunately, the specifications usually tie FDD systems to using cavity filters or 4-pole ceramic filters. Cavity filters run in the order of $35 each while ceramic filters can be in the $8 range. Most licensed bands do not have one standard structure but are flexible; that is, the TX and RX could be swapped in different geographical regions. This results in having to design several flavors of the filters, something that does not lend itself to mass production of the filters.

To give an idea of the filter requirements in FDD:

$$\text{Filter\_rej (dB)} = \text{Po (dBm/Hz)} - \text{Mask (dBc)} - [-174 + \text{NF-cochannel\_rej}]$$

For example, if power output Po $= -33$ dBm/Hz, in a 1 MHz signal bandwidth, output power is $+27$ dBm.

Mask of TX is = 60 dBc; that is, the thermal floor of TX is 60 dB below the Po.

NF is noise figure of Receiver = 5 dB.

Cochannel_rej is, how far in dB is the undesired signal below the desired signal, equal to 10 dB, that is, the undesired signal is 10 dB below the desired signal.

We get Filter_rej at the RX frequency of 86 dB. If the RX is 100 MHz away from the TX, this filter is an expensive cavity filter.

The full duplex nature of the circuit requires a separate TX and RX synthesizer. The RFIC die area is significantly impacted by the inductor of a resonant circuit; this is part of a voltage controlled oscillator (VCO) which is used in the synthesizer. Thus, two of these have a large impact on the cost of the RFIC.

A final note on FDD systems is that they are power hungry; this also increases the cost of the power system. Thus FDD is not an ideal platform to build portable or mobile radios.

FDD systems are typically deployed in licensed bands, for example, 5.8, 3.5, and 2.5 GHz: the spectrum is expensive. The cost of the spectrum forces the carriers to serve as many users as possible. Capacity must be optimized, which results in carriers favoring FDD architecture. Clearly, it is very desirable to have the Base Station work in FDD, but to reduce costs, the Subscriber Station could be a HFDD structure.

4.3.3 HFDD

Figure 4.3 shows a HFDD radio.

The HFDD architecture combines the benefits of the TDD systems while still trying to allow for frequency duplexing. The Base Station can operate in FDD and retain its capacity advantage over TDD

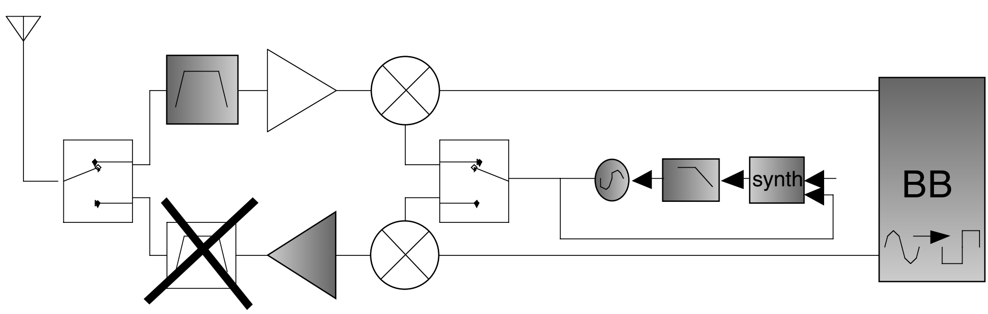


FIGURE 4.3 HFDD radio.

systems. This can lower the cost of the radio significantly at the Subscriber Station where the unit cost must be driven down.

The cost reduction appears in the form of relief in the RF TX filter, and since there is one synthesizer the die area of the RFIC shrinks. Power savings are also realized as in TDD systems.

Once again the collocation issues have to be addressed carefully. Self jamming is not a problem as in TDD but then too much relief on the TX filter can result in interference between users.

There is also a capacity loss at the Subscriber Station since the radio cannot simultaneously Transmit and Receive.

The HFDD structure can be used in both licensed and unlicensed bands. The Transmit and Receive can be at the same frequency as in TDD systems or separated by a frequency gap as in FDD. This type of radio is very flexible. Its cost structure approaches that of a TDD radio.

In summarizing the duplexing schemes, Intel's BB chip Rosedale can support both TDD and HFDD modes. This takes care of most of the Subscriber Stations. In a typical deployment the ratio between the Base Station and Subscriber Stations is 1–100, owing to the low volume of the Base Station. The physical (PHY) and media access control (MAC) layer need not be designed as a custom chip; a field programmable gate array (FPGA) could be cost effective. It is possible to connect two Rosedale BB chips together to support an FDD scenario for the Base Station.

We discuss various radio architectures in the following sections; these include IF- and I/Q-based architectures and some variants on these. Some of the interface between the radio and the BB chip is deliberated.

4.3.4 RF Interface

The BB chip digitizes the analog signal and performs signal processing. This PHY layer chip contains the blocks for filtering, automatic gain control (AGC), demodulation of data, security, and framing of data. The algorithms that do power measurements, such as AGC and RF selection can be taken care of by the lower-level MAC. As can be seen, there are common parameters such as AGC that are shared across the PHY, MAC, and radio.

The major blocks within a radio that need control from the baseband IC are AGC, frequency selection, sequencing of the TX/RX chain, monitoring of TX power, and any calibration functions, for example, I/Q imbalance. Each of these blocks are tightly coupled with the PHY and/or lower-level MAC.

A reasonable way to communicate with the radio is through a serial peripheral interface (SPI); it minimizes pins on the RFIC.

Usually the SPI is used to control the synthesizer. In order to make the interface more useful so that it can control the digital AGC of an RFIC and help perform measurements of power and temperature, the SPI needs to be a dedicated time-critical element. In this way, the SPI can respond to AGC, measurements, and frequency commands in a timely and predictable manner. A note of caution, however: traffic on the SPI could cause interference to the incoming signal and put spurs on the TX signal. Therefore, all SPI communication should only occur in the TX to RX time gaps. Other interface blocks are general purpose input/output (GPIO), pulse width modulators (PWM), DACs, and ADCs.

The AGC is split into RX AGC and TX AGC. In the RX AGC, response times may have to be rapid to cope with the changing RF channel in a mobile environment, in the order of μ sec. However in a fixed wireless application, the channel change is in the order of msec. The TX AGC can be relatively slow in steady state. However, in powering up the TX, the AGC may need to attain the correct power level in the μ sec time frame. Typically, the AGC is controlled through single-bit digital-to-analog (DA) converter, that is, sigma delta converters. Either of these methods have clock noise that needs to be filtered out. The trade-off here is that for a large slope of the RF AGC, the clock noise must be filtered to avoid distorting the signal. However, the filtering introduces a delay that slows down the AGC response. To increase the time response of the AGC, multibit DACs can be used.

The selection of the RF is performed through the SPI. For HFDD systems there is a settling time from TX to RX frequency, and the loading of the SPI is part of the timing budget.

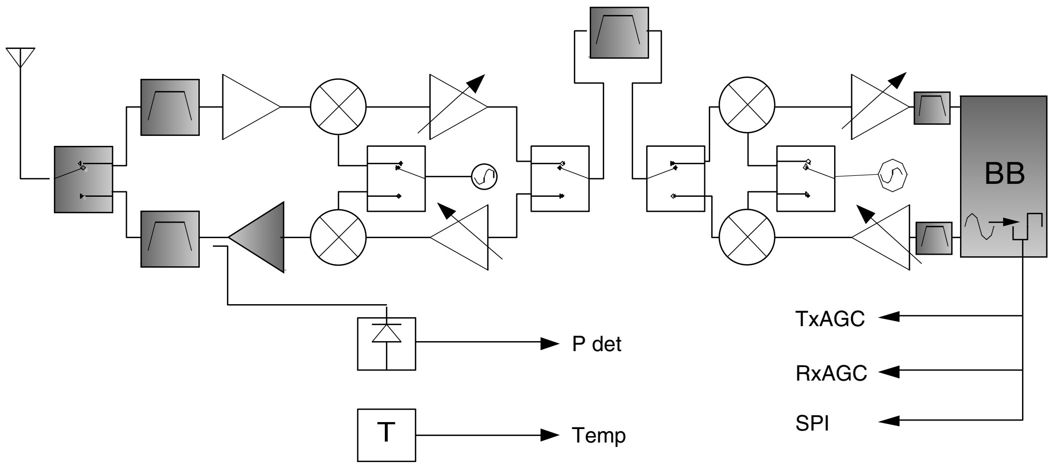


FIGURE 4.4 HFDD architecture.

Monitoring the temperature of the radio is a slow process; however, power measurements either from TX or RX require synchronization with the TX/RX timing gaps. Interfacing to the radio must take into account the sequencing of the radio; for example, in the case of the TX we need to switch the antenna, enable the TX and load frequency, change the TX Gain, turn on the PA, and finally ramp the modulation. Switching to the RX requires sequencing the TX down to avoid spurious emissions.

Two fundamental parameters drive radio design: noise and linearity. The goal is to attain as much dynamic range in the presence of undesired signals. This requires a distribution of gain and filtering through the TX or RX chain. Many architecture designers struggle with the placement of this gain and filtering. We look at some of these radio architectures in the next sections.

4.3.5 HFDD Architecture

The details of the HFDD architecture are shown in Figure 4.4.

There is a frequency separation between the TX and RX, so separate filtering is necessary in the RF front-end. However, the IF is shared between the TX and RX. A surface acoustic wave (SAW) filter provides for excellent adjacent/alternate channel rejection. There is a final frequency conversion to a lower IF that can be handled by an ADC. Much of the AGC range is at the lower IF. An AGC range of 70 dB is required; the absolute gain is higher to overcome losses. For the TX AGC, a 50 dB range is required. The AGC can be controlled through PWMs for analog AGC or GPIO for step attenuators.

Two synthesizers are necessary for the double conversion. The low-frequency synthesizer is fixed and does not have to be switched during the RX to TX change. The high-frequency synthesizer is the challenging block; it is required to settle within 100 μsec. The step size could be as low as 125 kHz in the 3.5 GHz band.

Several signals are also sent to the Baseband IC: TX power level (sometimes RX power level), temperature, and synthesizer lock detect. The power level is most important since power output has to be as close as possible to the intended value and still within regulations.

4.3.6 TDD Architecture

TDD is a good example of direct conversion transceivers or ZIF. Figure 4.5 is a block diagram of the ZIF architecture.

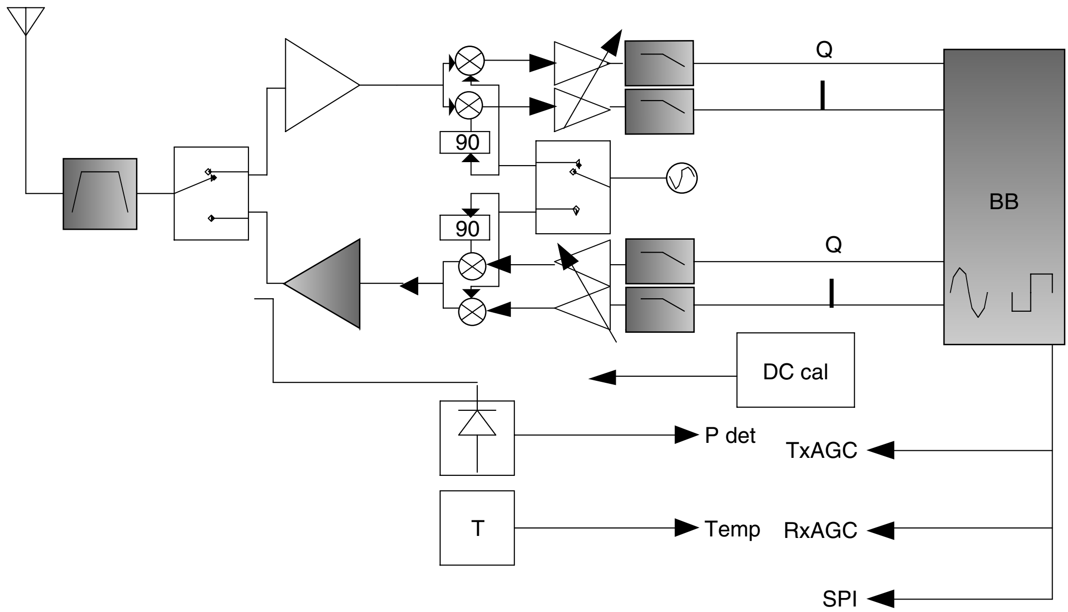


FIGURE 4.5 Block diagram of ZIF architecture.

The TX and RX frequencies are the same; hence, the RF filter can be shared. The down-conversion process is carried out with I/Q mixers; these consume a small area on the die. The issue with such mixers is that they need to be matched; otherwise, distortion is introduced. Also LO feedthrough effects tend to increase due to DC imbalances. These effects are significant since most of the gain is at the final conversion. The DC offset results in a reduction in the dynamic range of the AD since extra bits are required for this offset. A DC calibration circuit can be implemented to reduce the effect. In addition, I/Q imbalance will result in distortion. The problems are aggravated by temperature, gain changes, and frequency. By going to DC, low-pass filters can be used that are selective to channels. These can be implemented on chip and can save on cost. It must be noted that the on-chip low-pass filters do consume a large die area. They can also introduce noise. WiMAX has variable bandwidths ranging from 1 to 14 MHz but as the cutoff frequency is reduced there are significant challenges in the on-chip filter. For such ZIF schemes there must be an automatic frequency control (AFC) loop whereby the Baseband IC controls the reference oscillator of the RFIC. This ensures that any DC leakage terms stay at DC and do not spill over into the desired tones of the OFDM waveform.

4.3.7 I/Q Baseband Architecture 1

A variant of the HFDD and TDD architectures mentioned above is a combination shown in Figure 4.6.

This structure has the advantage that some filtering is carried out at an IF removing some of the strain on the DC filters.

In addition, power can be saved by having the final stage operate at lower frequencies. The issues related to I/Q mismatch and DC leakage are lessened by having less gain at DC and operating the mixers at an IF instead of an RF. Savings can be realized at the TX filtering: because the SAW can do most of the filtering there is no need for the TX low-pass filters. This has the added advantage that the I/Q mismatch from the low-pass filters is removed. One drawback is that two DA converters and two analog-to-digital (AD) converters are required.

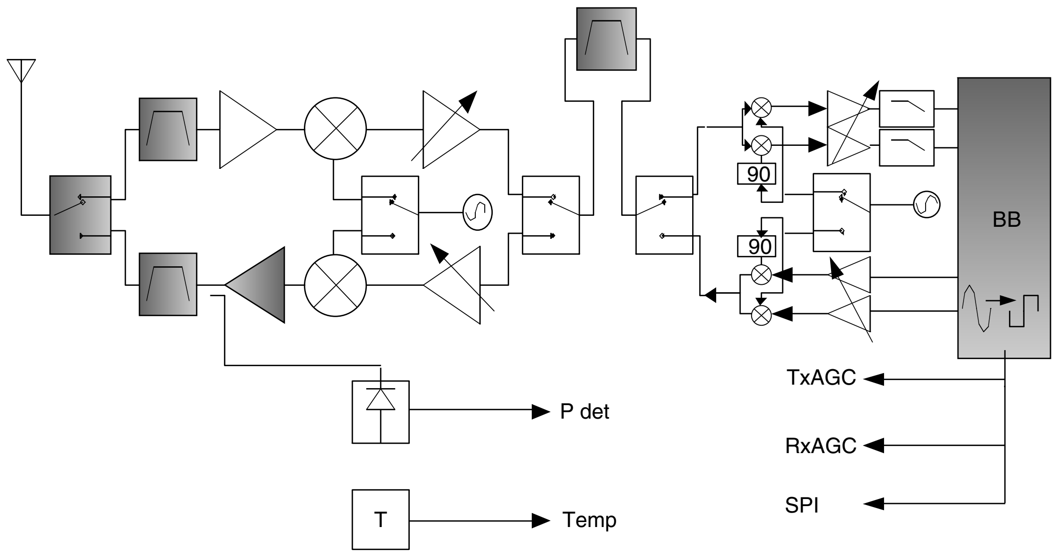


FIGURE 4.6 I/Q baseband architecture 1.

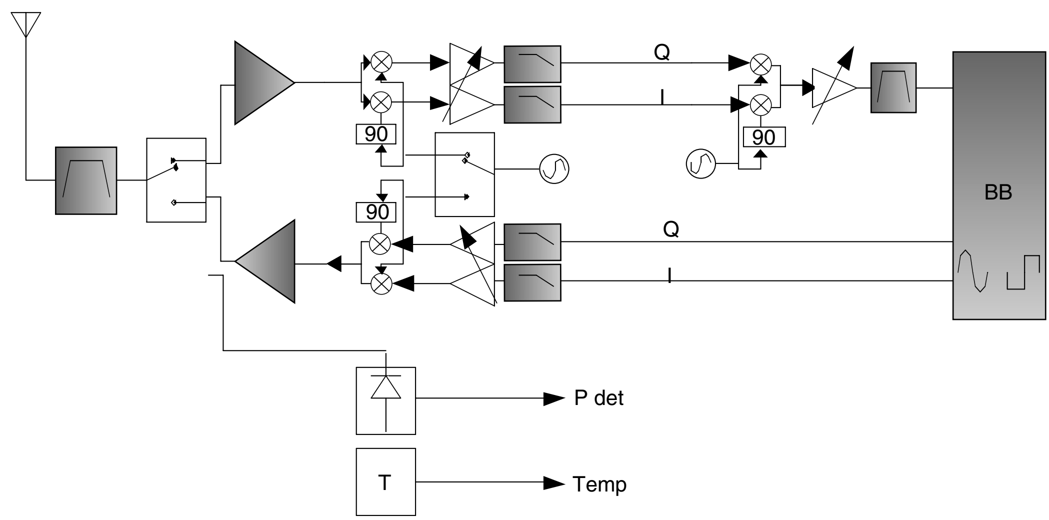


FIGURE 4.7 I/Q baseband architecture 2.

4.3.8 I/Q Baseband Architecture 2

To address the problems of the I/Q baseband radios another architecture is considered. Figure 4.7 shows an RX where the signal is mixed to DC then mixed up to a near zero IF (NZIF).

By going to DC the IF filter is removed and filtering can be done on-chip. To avoid DC and I/Q problems the signal is mixed to an IF. The choice of IF is greater than half the channel bandwidth. This structure allows the gain to be distributed between the DC and IF stages. In addition, as an added benefit, only one AD is required. For the TX stage, I/Q up-conversion is used.

4.3.9 RF Challenges for MIMO, AAS, and OFDMA

Antenna diversity is an important technique that can inexpensively enhance the performance of low-cost subscriber stations. It can help mitigate the effects of channel impairments like multipath, shadowing,

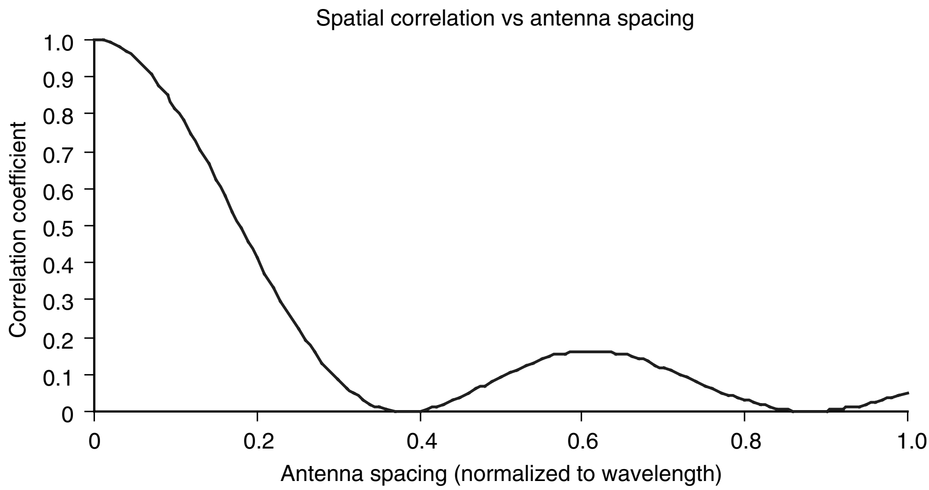


FIGURE 4.8 Bessel function approximation of the spatial correlation coefficient.

and interference that severely degrade a system's performance, and in some cases make it inoperable. By using multiple antennas, a system's link budget can be significantly improved by reducing channel fading and in some implementations by providing array gain. There are several designs, all of which yield excellent gains, that can be implemented, ranging from low to high complexity. The basic designs are selection diversity combining (SDC), equal gain combining (EGC), and maximum ratio combining (MRC). SDC is a scheme of sampling the receive performance of multiple antenna branches and selecting the branch that maximizes the receiver signal-to-noise ratio. To work properly each antenna branch must have relatively independent channel fading characteristics. To achieve this, the antennas are either spatially separated, use different polarization, or are a combination of both. The spatial correlation of antennas can be approximated by the zero order Bessel function given by the equation $\rho = J_0^2(2\pi d/\lambda)$ and shown in Figure 4.8. From Figure 4.8, it is seen that relatively uncorrelated antenna branches can be achieved for spatial separations greater than one-third a wavelength, supporting the requirement for small form-factor subscriber stations.

For optimal SDC performance the selection process and data gathering must be completed within the coherence time. The coherence time is the period over which a propagating wave preserves a near-constant phase relationship both temporally and spatially. After the coherence time has elapsed, the antennas should be resampled to account for expected channel variations and to allow for reselection of the optimal antenna. For a TDD system, where reciprocal uplink (UL) and downlink (DL) channel characteristics are expected, the selected receive antenna can also be used as the Transmit antenna. Although the SDC technique sounds rather simple, surprisingly large system gain improvements are possible if the algorithms can be designed effectively.

There are two figures of merit for judging the gain enhancement of an antenna diversity scheme. These are diversity gain and array gain. Under changing channel conditions, diversity gain is equivalent to the decrease in gain variance of local signal strength fluctuations of a multiantenna array system when compared with a single-antenna array system. The result of increased diversity gain is the reduction in fading depth. This is due to each antenna of a multiantenna system experiencing independent fading channels over frequency and time. The second figure-of-merit, array gain, is the accumulation of antenna gain associated with increased directivity via a multiantenna array system. In a typical system, as the number of antenna array elements grows, the gain increases 10*log (n), where n is the number of antenna array elements. This means a doubling of gain for every doubling of antenna elements.

TABLE 4.1 Performance Enhancement of Antenna Diversity

Antenna Diversity Scheme (4 Antenna Branches)	Antenna Gain (SUI3,SUI4 Model w/100 μ sec Rayleigh Delay Spread)	Implementation Complexity
Selection diversity combining	8 dB	Low
Equal gain combining (analog)	9 dB	Mid
Maximum ratio combining (analog)	10 dB	High
Maximum ratio combining (digital)	14 dB	High

The SDC scheme exhibits no array gain, as only one from n antennas is used at any instance. However, through spatial or polarization diversity, the SDC achieves stellar diversity gain, as shown in Table 4.1.

Another basic antenna diversity technique using multiple antennas is EGC. Instead of selecting one from n antennas, as in SDC, the algorithm combines the power of all antennas. The multiple independent signal branches are co-phased, the gain of each branch set to unity (equal gain), and then all branches combined. The EGC antenna diversity technique achieves diversity gain, while also producing array gain. Thus, EGC provides higher antenna diversity gain then SDC, as can be seen in Table 4.1. To achieve an antenna diversity benefit closer to optimal, MRC of the antenna elements can be used. This technique is similar to EGC, with the exception that the algorithm tries to optimally adjust both the phase and gain of each element prior to combining the power of all antennas. The summation of the signals may be done in either the analog or digital domain. When summation occurs in the digital domain, RF hardware for each independent antenna branch is required from RF to BB. When MRC is realized in the analog domain, summation may occur directly at RF. Performance is better when processing is done in the digital domain, as frequency selective channel characteristics are compensated for in each branch. In an analog MRC, only the average channel distortion over frequency is used to compensate for the amplitude and phase variation between array elements. In digital MRC, discrete frequency components across the signal bandwidth are co-phased and individually weighted based on SNR at the receiver. MRC realizes the highest antenna diversity gain compared with the other techniques discussed (refer to Table 4.1). Although the complexity is high, MRC implementation costs are decreasing through better RF integration and reduced CMOS geometries of the BB processor integrated circuit.

MIMO and AAS systems are used to improve link margins. Using MIMO requires multiple RF chains with multiple ADs. With integration, the cost of these multiple chains should come down. Isolation between the receive chains needs to be in the order of 20 dB, which is easy to accomplish. There are no matching requirements for the gain and phase between the RX chains, which means that the radio design is simplified. MIMO works well in TDD or FDD, and its improvements to link margins are observed in multipath environments.

In contrast, for AAS or beamforming systems, the TX and RX chain need to be matched across frequency and over gain and phase. However, the subscriber station does not have multiple chains. Such systems work well in TDD mode since the TX frequency is the same as RX frequency. Adaptive antenna systems (AAS) estimate the TX channel based on information they get from the RX channel, so having the same frequency improves these estimates.

OFDMA allows the RF channel to be split into subchannels. As a result, the power can be boosted since fewer tones are used. For users that do not TX much data on the UL, a smaller bandwidth can be allocated. Thus, more efficient use of the bandwidth can be made on a per-user basis. This technique does pose some challenges for the radio. Interference and noise between subchannels must be carefully considered over the whole transmit gain range. This problem is similar to the FDD case except there is no frequency separation. Therefore, noise performance and linearity must be excellent since there is no help from filtering. Another issue with OFDMA is that the RF must be maintained to <1% accuracy; otherwise, different users will collide with each other within the subchannels.

We have discussed various duplex schemes: RF architectures were outlined and some methods to improve link margin considered. Next, we discuss the particular circuit blocks within the RF system that are particular cost drivers.

4.4 RF System Blocks

There are three main areas of cost for a radio: synthesizer, PA, and filter.

4.4.1 Synthesizer

The synthesizer generates the LO that mixes with the incoming RF to create a lower frequency signal that can be digitized and processed by the Baseband IC. The WiMAX specifications call for a high-performance synthesizer. The synthesizer block takes up a large part of the RFIC die area and is therefore a costly component of the RFIC. The Integrated Phase Noise is <1 deg rms with an integration frequency of 1/20 of the tone spacing to half the channel bandwidth. Thus for the smaller bandwidths of 1.75 MHz, the integration of the phase noise can start as low as 100 Hz. For HFDD architectures, the TX to RX frequency has to settle within 100 μsec. The step size of the channel is 125 kHz in the 3.5 GHz band. To settle and maintain this step size, fractional synthesizers must be considered. It must be noted that as RF increases, obtaining phase noise <1 deg rms becomes a challenge. As well as all the radio LOs, the clock for the AD must be also viewed as an LO that adds phase noise to the overall jitter specification.

4.4.2 Power Amplifier

Wide-band digital modulation requires a high degree of linearity. Linearity implies higher power consumption. The trade-off between efficiency and linearity is a constant battle. For WiMAX, a PA can work at 4–5% efficiency for about a 6 dB backoff from output P1dB. Such a backoff results in about a 2.5% error vector magnitude (EVM) or 32 dBc of SNDR. With a class AB PA the efficiencies can run as high as 15–18% with similar EVM numbers.

A much overlooked parameter in PA design is settling time. When a PA is switched on from cold the power level will overshoot (or undershoot), then settle out. This settling time can be as poor as hundreds of milliseconds to get within 0.1 dB of the final value. For OFDM symbols, the RX has to estimate the power of a tone from the beginning of a frame to the end of a frame. If there is a droop of power from the beginning to the end of >0.1 dB across the frame, the BER for 64 quadrature amplitude modulation (QAM) will increase. The primary cause for this power droop is that the bias circuits and the output power field effect transistor (FET) are at thermally different points. Since this phenomenon is thermal, the effect can last for hundreds of milliseconds. To mitigate power droop, the bias circuits have to be placed as close to the output FETs as possible so they see the same temperature. In some cases, the PA may have to be turned on ahead of the TX cycle to allow the PA to stabilize and remove some of the droop. This implies having a trigger signal based on when data are to be transmitted. Having the MAC and PHY realize this trigger is not a simple matter. The budget of 100 μsec for HFDD is taken up by the synthesizer settling and any PA turn-on issues. A possible solution is to design the PA so that the PA settling is <5 μsec.

4.4.3 Filtering

Filtering is required to eliminate undesired signals from adjacent or alternate channels. Any noise from these immediate signals can leak noise into the desired band. Filtering at the receiver does not help; only a clean transmitted signal will prevent such degradation. Regulatory bodies control the transmitted mask.

For the adjacent channel problem the challenge is between linearity and filtering complexity. If the undesired channels are filtered out then less backoff in the radio is required and more of the AD bits are available for fading margin. SAW filters have depreciated in cost and are now in the <\$2 range for high volume. SAWs provide the optimum filtering. A significant drawback is that the technology does fix the maximum channel bandwidth that can be supported. Another issue is that it is difficult to support a large array of RF bands with a fixed IF. For spurious analysis, the optimum IF depends on the RF.

Filtering on-chip requires a large die area and as the channel bandwidth is reduced the die size increases. On-chip filters also produce more noise. A benefit is that the filter can be adjusted to accommodate the various bandwidths.

For I/Q-based designs, on-chip filters are necessary. The filters can be matched much more closely if on-chip. This minimizes the I/Q mismatch due to filtering. The final channel selectivity is performed in the Baseband IC using digital filters.

Filtering, like gain, must be distributed between the RF and subsequent down-conversions. The RF filtering is used to reduce the image and far blockers; that is, out of the RF band. The RF front-end must be linear enough to support the largest in-band blocker. In addition, reciprocal mixing of the LO with the undesired signal must be considered. The RF filters are typically >50 MHz wide and are constructed from various technologies each with different Q's. The larger the Q, the larger the size and the better filter shape. In FDD systems cavity filters may have to be used; these are large mechanical cavities and can cost >\$30 in high volume.

4.4.4 WiMax Specifications

We highlight some of the WiMAX RF specifications and contrast them with 802.11 specifications where possible. The specifications are broken into RX in Table 4.2 and TX in Table 4.3. It should be noted that most designs aim to do better than the standards; hence, these numbers should be viewed as the minimum requirements. In addition, we note the impact on the RFIC due to these specifications.

TABLE 4.2 RX Specifications

Parameter	802.11	WiMAX	Impact on RFIC
NF (dB)	10	7	The implication for the RFIC is that it may require an external LNA to meet a 5 dB NF
SNDR-64QAM (dBc)	<29	29	The implication for the RFIC is excellent phase noise for tone spacing of 5 kHz and linearity. For 802.11 the tone spacing is larger; that is, 300 kHz thus phase-noise requirement is less stringent
Alternate channel rejection (dBc)	NA	30	The AD bits may be used for allowing the adjacent channel through and some of the alternate channel. The digital filter would perform the bulk of the close-in channel filtering. Results in increase in linearity for RFIC
HFDD mode	No	Yes	More complicated synthesizer to support dual frequency
Channel BW (MHz)	10; 20	1.25; 1.75; 3.5; 7; 14; 5; 10; 20	The implication for the RFIC is that the smaller bandwidths result in a complicated synthesizer due to the smaller step size. Filtering for an array of bandwidths introduces adjacent channel compromises

TABLE 4.3 TX Specifications

Parameter	802.11	WiMAX	Impact on RFIC
Licensed band operation	No	Yes	The implication for RFIC is that the regulations are tighter and increase cost
AGC range (dB)	NA	50	The implication for RFIC is that linearity must be maintained over AGC range for 64 QAM
SNDR (dBc)	<31	31	The implication for RFIC is NF of TX chain, linearity, and phase noise
OFDMA	No	Yes	Noise and linearity must be maintained over the AGC range for in-channel cases
Smart antenna	No	Yes-Option	More RF chains for MIMO or matched RF chains for beamforming
Power output (dBm)	Restricted in unlicensed bands	<24 dBm	The implication for RFIC is PAs require higher efficiency, or even smart PA technology

4.5 Summary

WiMAX poses significant challenges to the RF subsystem. Several RF architectures were discussed in FDD, HFDD, and TDD modes. The cost-performance trade-offs in the various architectures were deliberated: these included IF- and BB-type radios. Some of the more important RF system blocks, synthesizers, PAs, and filtering that relate cost and specifications were discussed. Finally, some of the WiMAX radio specifications were highlighted and contrasted with 802.11, and the impact to RFIC development was noted.

4.6 Glossary

802.11 A standard for communication of wireless devices over reduced distances and reliability than the 802.16d/e standards.

802.16d A standard for fixed wireless, it has progressed to 802.16e a standard that allows for full mobility.

AAS Adpative antenna system, wireless systems that include beamforming using multiple antennas.

ADC Analog-to-digital converter, converts analog signals to digital.

AGC Automatic gain control, control circuitry or software used to maintain a constant level at the analog and digital divide. AGC is used to maximize the dynamic range of the radio.

BB Baseband, the processor that converts the received analog signal from the radio into digital and processes it and also converts the transmit digital stream into an analog domain for the radio.

BER Bit error rate, allows wireless systems to be compared at the PHY level.

DAC Digital-to-analog converter, converts digital signals to analog.

EGC Equal gain combining, a method of antenna diversity.

EVM Error vector magnitude, a measure of the fidelity of the signal.

FDD Frequency division duplex, where Transmit and Receive information is allocated on a frequency domain basis, the same time slot is used for both Transmit and Receive information; however, the Transmit and Receive frequencies are different.

FPGA Field programmable gate array, an integrated circuit that allows digital circuits to be reprogrammed.

GPIO General Purpose Input/Output, digital outputs used for general usage when communicating with the radio (e.g., turn on/off of the PA).

HFDD Half Frequency Division Duplex, where the Transmit and Receive information is allocated on a frequency domain basis, either the radio is in transmit mode or in receive mode; in addition, the Transmit and Receive frequencies are different.

I/Q Baseband frequency signals are signals that are centered at 0 Hz.

IF Intermediate frequency signal is one which is at a frequency higher than 0 Hz.

LO Local Oscillator A sub component of a radio that downconverts the radio frequency (RF) to an IF or 0 Hz through at least one mixer, conversely used to upconvert an IF or 0 Hz signal to RF.

LOS Line of sight, where the Antenna view between the Base station and the Subscriber is unobstructed.

MAC Medium access control, a layer of software that sits on top of the physical layer (PHY). It serves to control transmit power, framing, synchronization of packets, network entry, and quality of service.

MIMO Multiple In Multiple Out, advanced communication system that improves throughput and reliability of a wireless system it exploits multiple transmitters and receivers.

MRC Maximim ratio combining, a method of antenna diversity.

NZIF Near zero intermediate frequency, a very low IF.

OFDMA Orthogonal Frequency Division Multiplexing Access, allows for bandwidth allocation across multiple users. In WiMax, scaleable OFDMA is used and that allows the carrier spacing to remain constant as channel bandwidths and FFT sizes are changed.

P1dB Compression of the gain of an amplifier.

PA Power amplifier, the final amplifier in the transmit chain that delivers power to the antenna.

PWM Pulse width modulators, low-cost method for producing analog signals with a GPIO that can be used to affect radio circuits like amplifier gain.

Q Quality factor, of a component is the measure of inverse resistive loss in the material.

RFIC Radio frequency integrated circuit, a silicon chip or set of chips that form part of the radio. RFICs will typically contain amplifiers, mixers, synthesizers, and calibration circuits.

RX Receiver, the radio chain that downconverts from RF to a low frequency; in addition, it amplifies the received signal so that the baseband can process it. Typical receive signal range from 1 μ Vrms to 22 mVrms.

SAW Surface acoustic wave, a filter that can provide excellent characteristics in a small-form factor that cannot be achieved in an RFIC.

SDC Selection diversity combining, a method of antenna diversity.

SISO Single In Single Out, Wireless systems with a Single-Antenna input and a Single-Antenna output.

SNDR Signal to Noise plus Distortion, a measure of the fidelity of the signal.

SNR Signal-to-noise ratio, a measure of the fidelity of the signal.

SPI Serial peripheral interface, a three wire interface to control the radio.

TDD Time division duplex, where Transmit and Receive information is allocated on a time-domain basis, the same frequency is used for both Transmit and Receive information.

TX Transmitter, the radio chain that upconverts from low frequency to RF and delivers power to the antenna. Typical transmit chains output −30 to 30 dBm.

VCO Voltage controlled oscillator, a component used in the synthesizer subsystem of a radio used for up-conversion or down-conversion.

WiMax An organization that ensures interoperability between equipment vendors.

ZIF Zero IF, where the final frequency that is digitized by the baseband is at 0 Hz, that is, another name for I/Q radios.

References

A. Abidi, Direct Conversion Radio Trancievers for digital Communications, *IEEE Journal of Solid State-Circuits*, Vol. 30, No. 12, December 1995.

J. Crols and M. S. J. Steyaert, A 1.5 GHz highly Linear CMOS Downconversion, *IEEE Journal of Solid State-Circuits*, Vol. 30, No. 7, July 1995.

P. Gray and R. Meyer, *Analysis and Design of Analog Integrated Circuits, Third Edition*, John Wiley & Sons, Hoboken, New Jersey, 1993.

A. Parissinen, J. Jussila, J. Ryynanen, L. Sumanen, and K. A. I. Halonen, A 2-GHz Wide-Band Direct Conversion Reciever for WCDMA Applications, *IEEE Journal of Solid State-Circuits*, Vol. 34, No. 12, December 2000.

P. Vizmuller, *RF Circuits and Systems*, Artech House, Norwood, MA, 1995.

5

Broadband Wireless Access: High Rate, Point to Multipoint, Fixed Antenna Systems

Brian Petry
Member, IEEE

Broadband wireless access (BWA) broadly applies to systems providing radio communications access to a core network. Access is the key word because a BWA system by itself does not form a complete network, but only the access part. As the "last mile" between core networks and customers, BWA provides access services for a wide range of customers (also called subscribers), from homes to large enterprises. For enterprises such as small to large businesses, BWA supports such core networks as the public Internet, asynchronous transfer mode (ATM) networks, and the public switched telephone network (PSTN). Residential subscribers and small offices may not require access to such a broad set of core networks—Internet access is likely BWA's primary access function. BWA is meant to provide reliable, high throughput data services as an alternative to wired access technologies.

This chapter presents an overview of the requirements, functions, and protocols of BWA systems and describes some of today's efforts to standardize BWA interfaces.

5.1 Fundamental BWA Properties

Currently, the industry and standards committees are converging on a set of properties that BWA systems have, or should have, in common. A minimal BWA system consists of a single base station and a single subscriber station. The base station contains an interface, or interworking function (IWF), to a core network, and a radio "air" interface to the subscriber station. The subscriber station contains an interface to a customer premises network and of course, an air interface to the base station. Such a minimal system represents the point-to-point wireless transmission systems that have been in use for many years. Interesting BWA systems have more complex properties, the most central of which is point-to-multipoint

(P-MP) capability. A single base station can service multiple subscriber stations using the same radio channel. The P-MP property of BWA systems feature omnidirectional or shaped sector radio antennas at the base station that cover a geographic and spectral area that efficiently serves a set of customers given the allocation of radio spectrum. Multiple subscriber stations can receive the base station's downstream transmissions on the same radio channel. Depending on the density and data throughput requirements of subscribers in a given sector, multiple radio channels may be employed, thus overlaying sectors. The frequency bands used for BWA allow for conventional directional antennas. So, in the upstream transmission direction, a subscriber's radio antenna is usually highly directional, aimed at the base station. Such configuration of shaped sectors and directional antennas allow for flexible deployment of BWA systems and helps conserve radio spectrum by allowing frequency bands to be reused in nearby sectors.

With such P-MP functions and a sectorized approach, more BWA properties unfold and we find that BWA is similar to other other well-known access systems. A BWA deployment is cellular in nature, and like a cellular telephone deployment, requires complicated rules and guidelines that impact power transmission limits, frequency reuse, channel assignment, cell placement, and so forth. Also, since subscriber stations can share spectrum in both the upstream and downstream directions, yet do not communicate with each other using the air interface, BWA systems have properties very similar to hybrid fiber coaxial (HFC) access networks that coexist with cable television service. HFC networks also employ a base station (called a head end) and subscriber stations (called cable modems). And subscriber stations share channels in both downstream and upstream directions. Such HFC networks are now popularized by both proprietary systems and the Data-over-Cable System Interface Specifications (DOCSIS) industry standards [1]. In the downstream direction, digital video broadcast systems have properties similar to BWA. They employ base stations on the ground or onboard satellites: multiple subscribers tune their receivers to the same channels. With properties similar to cellular, cable modems, and digital video broadcast, BWA systems borrow many technical features from them.

5.2 BWA Fills Technology Gaps

Since BWA is access technology, it naturally competes with other broadband, high data rate access technologies, such as high data rate digital cellular service, digital subscriber line (DSL) on copper telephone wires, cable modems on coaxial TV cables, satellite-based access systems, and even optical access technologies on fiber or free space. To some, the application of BWA overlaps with these access technologies and also appears to fill in the gaps left by them. Following are some examples of technology overlaps where BWA fills in gaps.

High data rate digital cellular data service will be available by the time this book is published. This service is built "on top of" digital cellular telephone service. The maximum achievable data rate for these new "third generation" digital cellular systems is intended to be around 2.5 Mbps. At these maximum speeds, high data rate cellular competes with low-end BWA, but since BWA systems are not intended to be mobile, and utilize wider frequency bands, a BWA deployment should be able to offer higher data rates. Furthermore, a BWA service deployment does not require near ubiquitous service area coverage. Before service can be offered by mobile cellular services, service must be available throughout entire metropolitan areas. But for BWA, service can be offered where target customers are located before covering large areas. Thus, in addition to higher achievable data rates with BWA, the cost to reach the first subscribers should be much less.

Current DSL technology can reach out about 6 km from the telephone central office, but the achievable data rate degrades significantly as the maximum distance is reached. Currently, the maximum DSL data rate is around 8 Mbps. Asymmetric DSL (ADSL) provides higher data rates downstream than upstream, which is ideal for residential Internet access, but can be limiting for some business applications. BWA can fill in by providing much higher data rates further from telephone central offices. BWA protocols and deployment strategies enable the flexibility necessary to offer both asymmetric and symmetric services.

HFC cable modem technology, which is also asymmetric in nature, is ideal for residential subscribers. But many subscribers—potentially thousands—often share the same downstream channels and contend heavily for access to a limited number of available upstream channels. A key advantage of HFC is consistent channel characteristics throughout the network. With few exceptions, the fiber and coaxial cables deliver a consistent signal to subscribers over very long distances. BWA fills in, giving a service provider the flexibility to locate base stations and configure sectors to best service customers who need consistent, high data rates dedicated to them.

Satellite access systems are usually unidirectional, whereas less available bidirectional satellite-based service is more expensive. Either type of satellite access is asymmetric in nature: unidirectional service requires some sort of terrestrial "upstream," and many subscribers contend for the "uplink" in bidirectional access systems. Satellites in geostationary Earth orbits (GEO) impose a minimum transit delay of 240 ms on transmissions between ground stations. Before a satellite access system can be profitable, it must overcome the notable initial expense of launching satellites or leasing bandwidth on a limited number of existing satellites by registering many subscribers. Yet, satellite access services offer extremely wide geographic coverage with no infrastructure planning, which is especially attractive for rural or remote service areas that DSL and cable modems do not reach. Perhaps high data rate, global service coverage by low Earth orbiting (LEO) satellites will someday overcome some of GEO's limitations. BWA fills in by allowing service providers to locate base stations and infrastructure near subscribers that should be more cost effective and impose less delay than satellite services.

Optical access technologies offer unbeatable performance in data rate, reliability, and range, where access to fiber-optic cable is available. But in most areas, only large businesses have access to fiber. New technology to overcome this limitation, and avoid digging trenches and pulling fiber into the customer premises is free space optical, which employs lasers to extend between a business and a point where fiber is more readily accessible. Since BWA base stations could also be employed at fiber access points to reach non-fiber-capable subscribers, both BWA and free space optical require less infrastructure planning such as digging, tunneling, and pulling cables under streets. Although optical can offer an order of magnitude increase in data rate over the comparatively lower frequency/higher wavelength BWA radio communications, BWA can have an advantage in some instances because BWA typically has a longer range and its sector-based coverage allows multiple subscribers to be serviced by a single base station.

Given these gaps left by other broadband access technologies, even with directly overlapping competition in many areas, the long-term success of BWA technology is virtually ensured.

5.3 BWA Frequency Bands and Market Factors

Globally, a wide range of frequency bands are available for use by BWA systems. To date, systems that implement BWA fall into roughly two categories: those that operate at high frequencies (roughly 10–60 GHz) and those that operate at low frequencies (2–11 GHz). Systems in the low frequency category may be further subdivided into those that operate in licensed vs. unlicensed bands. Unlicensed low frequency bands are sometimes considered separately because of the variations of emitted power restrictions imposed by regulatory agencies and larger potential for interference by other "unlicensed" technologies. The high frequencies have significantly different characteristics than the low frequencies that impact the expense of equipment, base station locations, range of coverage, and other factors. The key differing characteristics in turn impact the type of subscriber and types of services offered as will be seen later in this article.

Even though available spectrum varies, most nationalities and regulatory bodies recognize the vicinity of 30 GHz, with wide bands typically available, for use by BWA. In the United States, for instance, the FCC has allocated local multipoint distribution service (LMDS) bands for BWA. That, coupled with the availability of radio experience, borrowed from military purposes and satellite communications, influenced the BWA industry to focus their efforts in this area. BWA in the vicinity of 30 GHz is thus also a target area for standardization of interoperable BWA systems. Available spectrum for lower frequencies, 2–11 GHz, varies widely by geography and regulatory body. In the United States, for instance, the FCC has allocated several bands called multipoint/multichannel distribution services (MDS) and licensed them for BWA use. The industry is also targeting the lower spectrum, both licensed and unlicensed, for standardization.

Radio communications around 30 GHz have some important implications for BWA. For subscriber stations, directional radio antennas are practical. For base stations, so are shaped sector antennas. But two key impairments limit how such BWA systems are deployed: line-of-site and rain. BWA at 30 GHz almost strictly requires a line-of-sight path to operate effectively. Even foliage can prohibitively absorb the radio energy. Some near line-of-sight scenarios, such as a radio beam that passes in close proximity to reflective surfaces like metal sheets or wet roofs, can also cause significant communications impairments. Rain can be penetrated, depending on the distance between subscriber and base station, the droplet size, and rate of precipitation. BWA service providers pay close attention to climate zones and historical weather data to plan deployments. In rainy areas where subscribers require high data rate services, small cell sizes can satisfy a particular guaranteed service availability. Also, to accommodate changing rain conditions, BWA systems offer adaptive transmit power control. As the rate of precipitation increases, transmit power is boosted as necessary. The base station and subscriber station coordinate with each other to boost or decrease transmit power.

Impairments aside, equipment cost is an important issue with 30 GHz BWA systems. As of today, of all the components in a BWA station, the radio power amplifier contributes most to system cost. Furthermore, since the subscriber antenna must be located outdoors (to overcome the aforementioned impairments), installation cost contributes to the equation. A subscriber installation consists of an indoor unit (IDU) that typically houses the digital equipment, modem, control functions, and interface to the subscriber network, and an outdoor unit (ODU), which typically houses the amplifier and antenna. Today these factors, combined with the aforementioned impairments, typically limit the use of 30 GHz BWA systems to businesses that both need the higher-end of achievable data rates and can afford the equipment. BWA technology achieves data rates delivered to a subscriber in a wide range, 2–155 Mbps. The cost of 30 GHz BWA technology may render the lower end of the range impractical. However, many people project the cost of 30 GHz BWA equipment to drop as the years go by, to the point where residential service will be practical.

In the lower spectrum for BWA systems, in the range of approximately 2–11 GHz, line-of-sight and rain are not as critical impairments. Here, a key issue to contend with is interference due to reflections, also called multipath. A receiver, either base station or subscriber, may have to cope with multiple copies of the signal, received at different delays, due to reflections off buildings or other large objects in a sector. So, different modulation techniques may be employed in these lower frequency BWA systems, as opposed to high frequency systems, to compensate for multipath. Furthermore, if the additional expense can be justified, subscribers and/or base stations, could employ spatial processing to combine the main signal with its reflections and thus find a stronger signal that has more data capacity than the main signal by itself. Such spatial processing requires at least two antennas and radio receivers. In some cases, it may even be beneficial for a base station to employ induced multipath, using multiple transmit antennas, perhaps aimed at reflective objects, to reach subscribers, even those hidden by obstructions, with a better combined signal than just one.

Unlike BWA near 30 GHz, BWA in the lower spectrum today has the advantage of less expensive equipment. Also, it may be feasible in some deployments for the subscriber antenna to be located indoors. Further, the achievable data rates are typically lower than at 30 GHz, with smaller channel bandwidths, in the range of about 2–15 Mbps. Although some promise 30 GHz equipment costs will drop, these factors make lower frequency BWA more attractive to residences and small businesses today.

Due to the differing requirements of businesses and residences and the different capabilities of higher frequency BWA versus lower, the types of service offered is naturally divided as well. Businesses will typically subscribe to BWA at the higher frequencies, around 30 GHz, and employ services that carry guaranteed quality of service for both data and voice communications. In the business category, multi-tenant office buildings and dwellings are also lumped in. At multi-tenant sites, multiple paying subscribers share one BWA radio and each subscriber may require different data or voice services. For data, Internet Protocol (IP) service is of prime importance, but large businesses also rely on wide area network technologies like asynchronous transfer mode (ATM) and frame relay that BWA must efficiently transport. To date, ATM's capabilities offer practical methods for dedicating, partitioning, and prioritizing data flows, generally called quality of service (QoS). But as time goes on (perhaps by this reading) IP-based QoS capabilities will overtake ATM. So, for both residential and business purposes, IP service will be the data service of choice in the future. Besides data, businesses rely on traditional telephony links to local telephone service providers. Business telephony services, for medium-to-large enterprises, utilize time division multiplexed (TDM) telephone circuits on copper wires to aggregate voice calls. Some BWA systems have the means to efficiently transport such aggregated voice circuits. Due to the economic and performance differences between low frequency BWA and high frequency BWA, low frequency BWA generally carries residential- and small business-oriented services, whereas high frequency BWA carries small- to large-enterprise services.

Since BWA equipment for the lower frequencies may be less expensive and less sensitive to radio directionality, and therefore more practical to cover large areas such as residential environments, subscriber equipment can potentially be nomadic. Nomadic means that the equipment may be moved quickly and easily from one location to another, but is not expected to be usable while in transit. Whereas at the higher frequencies, with more expensive subscriber equipment, the decoupling of indoor and outdoor units, the highly directional nature of radio communications in that range, and subscriber-oriented services provisioned at the base station, subscriber stations are fixed. Once they are installed, they do not move unless the subscriber terminates service and re-subscribes somewhere else.

5.4 Standards Activities

Several standards activities are under way to enable interoperability between vendors of BWA equipment. The standardization efforts are primarily focused on an interoperable "air interface" that defines how compliant base stations interoperate with compliant subscriber stations. By this reading, some of the standards may have been completed—the reader is encouraged to check the status of BWA standardization. Some standards groups archive contributions by industry participants and those archives, along with the actual published standards, provide important insights into BWA technology. Currently, most activity is centered around the Institute for Electronics and Electrical Engineers (IEEE) Local Area Network/ Metropolitan Area Network (LAN/MAN) Standards Committee (LMSC), which authors the IEEE 802 series of data network standards. Within LMSC, the 802.16 working group authors BWA standards. The other notable BWA standards effort, under the direction of the European Telecommunications Standards Institute (ETSI), is a project called Broadband Radio Access Networks/HyperAccess (BRAN/HA). The IEEE LMSC is an organization that has international membership and has the means to promote their standards to "international standard" status through the International Organization for Standardization (ISO) as does ETSI. But ETSI standards draw from a European base, whereas LMSC draws from a more international base of participation. Even so, the LMSC and BRAN/HA groups, although they strive to develop standards each with a different approach, have many common members who desire to promote a single, international standard. Hopefully, the reader will have discovered that the two groups have converged on one standard that enables internationally harmonized BWA interoperability.

To date, the IEEE 802.16 working group has segmented their activities into three main areas: BWA interoperability at bands around 30 GHz (802.16.1), a recommended practice for the coexistence of BWA systems (802.16.2) and BWA interoperability for licensed bands between 2 and 11 GHz (802.16.3). By the time this book is published, more standards activities may have been added, such as interoperability for some unlicensed bands. The ETSI BRAN/HA group is focused on interoperability in bands around 30 GHz.

Past standards activities were efforts to agree on how to adapt existing technologies for BWA: cable modems and digital video broadcast. A BWA air interface, as similar to DOCSIS cable modems as possible, was standardized by the radio sector of the International Telecommunications Union (ITU) under the ITU-R Joint Rappateur's Group (JRG) 9B committee [2]. The Digital Audio-Video Council (DAVIC) has standardized audio and video transport using techniques similar to BWA [3]. Similarly, the Digital Video

Broadcasting (DVB) industry consortium, noted for having published important standards for satellite digital video broadcast, has also published standards, through ETSI, for terrestrial-based digital television broadcast over both cable television networks and wireless. DVB has defined the means to broadcast digital video in both the "low" (<10 Gbps) and "high" (>10 Gbps) BWA spectra [4, 5]. These standards enabled interoperability of early BWA deployment by utilizing existing subsystems and components. Technology from them provided a basis for both the IEEE LMSC and ETSI BRAN/HA standardization processes. However, the current IEEE and ETSI efforts strive to define protocols with features and nuances more particular to efficient BWA communications.

5.5 Technical Issues: Interfaces and Protocols

A BWA access network is perhaps best described by its air interface: what goes on between the base station and subscriber stations. Other important interfaces exist in BWA systems, such as:

- Interfaces to core networks like ATM, Frame Relay, IP, and PSTN
- Interfaces to subscriber networks like ATM, Ethernet, Token Ring, and private branch exchange (PBX) telephone systems
- The interface between indoor unit (IDU) and outdoor unit (ODU)
- Interfaces to back-haul links, both wired and wireless, for remote base stations not co-located with core networks
- Air interface repeaters and reflectors

These other interfaces are outside the scope of this article. However, understanding their requirements is important to consider how a BWA air interface can best support external interfaces, particularly how the air interface supports their unique throughput, delay, and QoS requirements.

5.5.1 Protocols and Layering

Network subsystems following the IEEE LMSC reference model [6] focus on the lower two layers of the ISO Basic Reference Model for Open Systems Interconnection [7]. The air interface of a BWA system is also best described by these two layers. In LMSC standards, layers one and two, the physical and data link layers, are typically further subdivided. For BWA, the important subdivision of layer 2 is the medium access control (MAC) sublayer. This layer defines the protocols and procedures by which network nodes contend for access to a shared channel, or physical layer. In a BWA system, since frequency channels are shared among subscriber stations in both the downstream and upstream directions, MAC layer services are critical for efficient operation. The physical layer (PHY) of a BWA system is responsible for providing a raw communications channel, employing modulation and error correction technology appropriate for BWA.

Other critical functions, some of which may reside outside the MAC and PHY layers, must also be defined for an interoperable air interface: security and management. Security is divided two areas: a subscriber's authorized use of a base station and associated radio channels and privacy of transported data. Since the communications channel is wireless, it is subject to abuse by intruders, observers, and those seeking to deny service. BWA security protocols must be well defined to provide wire-like security and allow for interoperability. Since to a great extent, HFC cable TV access networks are very similar to BWA regarding security requirements, BWA borrows heavily from the security technology of such cable systems. Similarly, interoperable management mechanisms and protocols include the means to provision, control and monitor subscribers stations and base stations.

5.5.1.1 The Physical Layer

The PHY is designed with several fundamental goals in mind: spectral efficiency, reliability, and performance. However, these are not independent goals. We can not have the best of all three because each of those goals affects the others: too much of one means too little of the others. But reliability and performance levels are likely to be specified. And once they are specified, spectral efficiency can be somewhat optimized. One measure of reliability is the bit error ratio (BER), the ratio of the number of bit errors to the number of non-errored bits, delivered by a PHY receiver to the MAC layer. The physical layer must provide for better than 10^{-6} BER, and hopefully closer to 10^{-9}. The larger error ratio may only be suitable for some voice services, whereas a ratio closer to the lower end of the range is required for reliable data services that could offer equivalent error performance as LANs. Reliability is related to availability. Business subscribers often require contracts that guarantee a certain level of availability. For instance, a service provider may promise that the air interface be available to offer guaranteed reliability and performance 99.99% (also called "four nines") of the time.

Performance goals specify minimum data rates. Since, in BWA systems, the spectrum is shared by subscribers, and allocation of capacity among them is up to the MAC layer, the PHY is more concerned with the aggregate capacity of a single radio channel in one sector of a base station than for capacity to a given subscriber. But if one subscriber would offer to purchase all the available capacity, the service provider would undoubtedly comply. For instance, a capacity goal currently set by the BRAN/HA committee is 25 Mbps on a 28 MHz channel. Without considering deployment scenarios, however, PHY goals are meaningless. Obviously, higher capacity and reliability could be better achieved by shorter, narrower sectors (smaller cells) rather than wider, longer sectors (larger cells). And the same sized sector in a rainy, or obstructed, terrain offers less guaranteed capacity than one in the flattest part of the desert. In any case, the industry seems to be converging on a goal to provide at least 1 bps/Hz capacity in an approximately 25 MHz wide channel with a BER of 10^{-8}. Many deployments should be able to offer much greater capacity.

In addition to such fundamental goals, other factors affect the choice of PHY protocols and procedures. One is duplex mode. The duplex mode can affect the cost of equipment, and some regulatory bodies may limit the choice of duplex mode in certain bands. Three duplex modes are considered for BWA: frequency division duplex (FDD), time division duplex (TDD), and half-duplex FDD (H-FDD). In FDD, a radio channel is designated for either upstream- or downstream-only use. Some bands are regulated such that a channel could only be upstream or downstream, thus requiring FDD if such bands are to be utilized. In TDD mode, one channel is used for both upstream and downstream communications. TDD-capable BWA equipment thus ping-pongs between transmit and receive mode within a single channel; all equipment in a sector is synchronized to divisions between transmit and receive. TDD is useful for bands in which the number of available, or licensed, channels is limited. TDD also allows for asymmetric service without reconfiguring the bandwidth of FDD channels. For instance, a service provider may determine that a residential deployment is more apt to utilize more downstream bandwidth than upstream. Then, rather than reallocating or sub-channeling FDD channels, the service provider can designate more time in a channel for downstream communications than upstream. Additionally, TDD equipment could potentially be less expensive than FDD equipment since components may be shared between the upstream and downstream paths and the cost of a duplexor may be eliminated. However, the third option, H-FDD, is a reasonable compromise between TDD and FDD. In H-FDD mode, a subscriber station decides when it can transmit and when it can receive, but cannot receive while transmitting. But the base station is usually full duplex, or FDD. For subscriber stations, H-FDD equipment can achieve the same cost savings as TDD, and offers the flexibility of asymmetric service. But H-FDD does not require all subscribers in a sector to synchronize on the same allocated time between transmit and receive.

Another important factor affecting spectral efficiency, upgradability, and flexible deployment scenarios, is adaptive modulation. In BWA, the channel characteristics vary much more widely than wired access systems. Rain, interference and other factors can affect subscriber stations individually in a sector, whereas in wired networks, such as HFC cable TV, the channel characteristics are consistent. Thus, to make good use of available bandwidth in favorable channel conditions, subscribers that can take advantage of higher data rates should be allowed to do so. And when it rains in one portion of a sector, or other impairments such as interference occur, subscriber stations can adapt to the channel conditions by reducing the data rate (although transmit power level adjustment is usually the first adaptive tool

BWA stations use when it rains). Besides adapting to channel conditions, adaptive modulation facilitates future deployment of newer modulation techniques while retaining compatibility with currently installed subscriber stations. When the service provider upgrades a base station and offers better modulation to new customers, not all subscriber stations become obsolete. To achieve the most flexibility in adaptive modulation, BWA employs "per-subscriber" adaptive modulation to both downstream and upstream communications. Per-subscriber means that each subscriber station can communicate with the base station using a different modulation technique, within the same channel. Some BWA equipment offers per-subscriber adaptive modulation in both the downstream and upstream directions. But other equipment implements a compromise that allows for equipment or components, similar to cable modems or digital video broadcast systems, to require all subscribers to use the same modulation in the downstream direction at any one point in time. Most BWA equipment implements adaptive modulation in the upstream direction. The overriding factor for the PHY layer, with regard to adaptive modulation, is burst mode. Adaptive modulation generally requires burst mode communications at the PHY layer. Time is divided into small units in which stations transmit independent bursts of data. If the downstream employs per-subscriber adaptive modulation, the base station transmits independent bursts to the subscribers. Each burst contains enough information for the receiver to perform synchronization and equalization. However, if per-subscriber adaptive modulation is not employed in the downstream direction, the base station can transmit in continuous mode, in very large, continuous chunks, each chunk potentially containing data destined for multiple subscribers. In burst mode downstream communications, the base station informs subscriber stations, in advance, which burst is theirs. In this way, a subscriber station is not required to demodulate each burst to discover which bursts are for the station, but only to demodulate the "map." The base station encodes the map using the least common denominator modulation type so all subscriber stations can decode it. Conversely, continuous mode downstream, in which per-subscriber adaptive modulation is not used, requires all subscriber stations to demodulate prior to discovering which portions of data are destined for the station. So, per-subscriber adaptive modulation in the downstream affords more flexibility, but a continuous mode downstream may also be used. The standards efforts currently are attempting to work out how both downstream modes may be allowed and yet still have an interoperable standard.

Burst size and the choice of continuous downstream mode in turn affect the choice of error correction coding. Some coding schemes are more efficient with large block sizes, whereas others are more efficient with smaller block sizes.

The fundamental choice of modulation type for BWA varies between the upper BWA bands (~30 GHz) and lower bands (~2–11 GHz). In the upper bands, the industry seems to be converging on Quadrature Phase Shift Keying (QPSK) and various levels of Quadrature Amplitude Modulation (QAM). These techniques may also be used in the lower bands, but given the multipath effects that are much more prevalent in the lower bands, BWA equipment is likely to employ Orthogonal Frequency Division Multiplexing (OFDM) or Code Division Multiple Access (CDMA) technology that have inherent properties to mitigate the effects of multipath and spread transmit energy evenly throughout the channel spectrum.

5.5.1.2 The Medium Access Control Layer

The primary responsibility of the Medium Access Control Layer (MAC) is to allocate capacity among subscriber stations in a way that preserves quality-of-service (QoS) requirements of the services it transports. For instance, traditional telephony and video services could require a constant, dedicated capacity with fixed delay properties. But other data transport services could tolerate more bursty capacity allocations and a higher degree of delay variation. ATM service is notable for its QoS definitions [8]. Although not mature as of this writing, the IP QoS definitions are also notable [9, 10]. Though QoS-based capacity allocation is a complex process, the BWA MAC protocol defines the mechanisms to preserve QoS as it transports data. Yet the MAC protocol does not fully define *how* MAC mechanisms are to be used. At first glance, this does not seem to make sense, but it allows the MAC protocol to be defined in as simple terms as possible and leave it up to implementations of base stations and subscriber stations how to best utilize the mechanism that the protocol defines. This approach also allows BWA vendors to differentiate their equipment and still retain interoperability. To simplify capacity allocation, the smarts of QoS implementation reside in the base station, since it is a central point in a BWA sector and is in constant communication with all of the subscriber stations in a sector. The base station is also administered by the service provider, and therefore can serve as the best point of control to keep subscribers from exceeding their contractual capacity limitations and priorities.

Capacity Allocation Mechanisms—An overview of the mechanisms employed by the MAC layer to allocate capacity follows. In the downstream direction, the MAC protocol informs subscriber stations what data belongs to what subscriber by means of per-subscriber addressing and within a subscriber, by per-data-flow addressing. All subscribers in a sector "listen" to the downstream data flow and pick off transmissions belonging to them. If the downstream channel employs per-subscriber adaptive modulation, some subscriber stations may not be able to decode the modulation destined to other subscribers. In this case, the base station informs subscribers what bursts it should observe, with a downstream "map." The downstream map indicates what offsets in a subsequent transmission may contain data for the specified subscriber. The MAC must communicate this information to the PHY layer to control its demodulation.

For upstream capacity allocation and reservation, the MAC employs slightly more complicated schemes. The upstream channel is the central point of contention: all subscriber stations in a channel are contending for access to transmit in the upstream channel. Some subscribers require constant periodic access, others require bursty access with minimum and maximum reservation limits. Still other data flows may not require any long-standing reservations but can request a chunk of capacity when needed and survive the inherent access delay until the base station satisfies the request. On top of these varying data flow requirements, which are specified by subscriber stations and granted by the base station, priorities increase complications. The base station administers both priorities and QoS parameters of each data flow in each subscriber station. How a base station keeps track of all the flows of subscribers and how it actually meets the reservation requirements is usually beyond the scope of the BWA air interface in standards documents. But base stations likely employ well-known queuing algorithms and reservation lists to ensure that it assigns capacity fairly and meets subscribers' contractual obligations. Yet, as mentioned earlier, room is left for BWA base station vendors to employ proprietary "tricks" to differentiate their equipment from others. To communicate capacity allocation to subscribers, the base station divides time into multi-access frames (e.g., on the order of 1–5 ms) in which multiple subscribers are assigned capacity. To accomplish this, a fixed allocation unit, or time slot, is defined. So, the upstream channel is divided into small, fixed-length time slots (e.g., on the order of 10 μs) and the base station periodically transmits a "map" of slot assignments to all subscribers in a channel. The slot assignments inform the subscriber stations which slots are theirs for the upcoming multi-access frame.

Periodically, a set of upstream slots is reserved for "open contention." That is, any subscriber is authorized to transmit during an open contention period. A subscriber can utilize open contention for initial sign-on to the network (called "registration"), to transmit a request for upstream capacity, or even to transmit a small amount of data. Since a transmission may collide with that of another subscriber station, a collision avoidance scheme is used. A subscriber station initiates transmission in a randomly chosen open contention slot, but cannot immediately detect that its transmission collided with another. The only way a subscriber station can determine if its transmission collided is if it receives no acknowledgment from the base station. In this case, the subscriber backs off a random number of open contention slots before attempting another transmission. The process continues, with the random number range getting exponentially larger on each attempt, until the transmission succeeds. The random back-off interval is typically truncated at the sixteenth attempt, when the subscriber station starts over with its next attempt in the original random number range. This back-off scheme is called "truncated binary exponential back-off," and is employed by popular MAC protocols such as Ethernet [11].

To mitigate the effects of potentially excessive collisions during open contention, the MAC protocol defines a means to request bandwidth during assigned slots in which no collision would happen. For instance, active subscriber stations may receive from the base station a periodic slot for requesting capacity or requesting a change in a prior reservation. This form of allocation-for-a-reservation-request is called a "poll." Also, the MAC protocol provides a means to "piggy-back" a request for capacity with a normal upstream data transmission. Subscriber stations that have been inactive may receive less frequent polls from the base station so as to conserve bandwidth. So, with a means for contentionless bandwidth reservation, the only time subscriber stations need to use the open contention window is for initial registration.

Slot-based reservations require that the base stations and subscribers be synchronized. Of course, the base station provides a timing base for all subscriber stations. To achieve accurate timing, subscriber stations need to determine how far they are from the base station so their transmissions can be scheduled to reach the base station at the exact point in time, relative to each other. The procedure to determine this distance, which is not really a measured linear distance, but a measurement of time, is called "ranging." Each subscriber station, coordinating with the base station, performs ranging during its registration process.

To maintain efficient use of bandwidth and accommodate PHY requirements of transmit power control, and flexible duplex modes, the MAC protocol performs even more gyrations. If interested, the reader is encouraged to read BWA MAC protocol standards, or drafts in progress, to learn more.

5.5.1.3 Automatic Repeat Request (ARQ) Layer

Some BWA systems trade off the bandwidth normally consumed by the PHY's error correction coding for the potential delays of ARQ protocol. An ARQ protocol employs sequence numbering and retransmissions to provide a reliable air link between base station and subscriber. ARQ requires more buffering in both the base station and subscriber station than systems without ARQ. But even with a highly-coded PHY, some subscriber stations may be located in high interference or burst-noise environments in which error correction falls apart. In such situations, ARQ can maintain performance, or ensure the service meets contractual availability and reliability requirements. Standards groups seem to be converging on allowing the use of ARQ, but not requiring it. The MAC protocol is then specified so that when ARQ is not used, no additional overhead is allocated just to allow the ARQ option.

5.6 Conclusion

This article has provided an overview of how BWA fits in with other broadband access technologies. It was also a short primer on BWA protocols and standards. To learn more about BWA, the reader is encouraged to read currently available standards documents, various radio communications technical journals, and consult with vendors of BWA equipment.

References

1. SCTE SP-RFI-105-981010: *Data-Over-Cable Service Interface Specifications: Radio Frequency Interface Specification,* The Society of Cable Telecommunications Engineers, Exton, Pennsylvania, 1999.
2. Draft Recommendation F.9B/BWA—Radio transmission systems for fixed broadband wireless access (BWA) based on cable modem standard, International Telecommunications Union, Geneva, 1999.
3. DAVIC 1.4.1 *Specification Part 8: Lower Layer Protocols and Physical Interfaces,* Digital Audio-Visual Council, Geneva, 1999.
4. ETS 300 748, *Digital Video Broadcasting (DVB): Multipoint Video Distribution Systems (MVDS) at 10 GHz and above*, European Telecommunications Standards Institute, Geneva, 1997.
5. ETS 300 749, *Digital Video Broadcasting (DVB): Digital Video Broadcasting (DVB); Microwave Multipoint Distribution Systems (MMDS) below 10 GHz*, European Telecommunications Standards Institute, Geneva, 1997.

6. IEEE Std 802-1990, *IEEE Standards for Local and Metropolitan Area Networks: Overview and Architecture*, Institute for Electrical and Electronics Engineers, Piscataway, NJ, 1990.
7. ISO/IEC 7498-1:1994, *Information Technology—Open Systems Interconnection—Basic Reference Model: The Basic Model*, International Organization for Standardization, Geneva, 1994.
8. Bermejo, L. P. et al., Service characteristics and traffic models in broadband ISDN, *Electrical Commun.,* 64-2/3, 132–138, 1990.
9. Blake, S. et al., *RFC-2475 An Architecture for Differentiated Service*, Internet Engineering Task Force, 1998.
10. Braden, R. et al., *RFC-2205 Resource Reservation Protocol (RSVP)—Version 1 Functional Specification*, Internet Engineering Task Force, 1997.
11. IEEE Std 802.3, *Information Technology—Telecommunications and Information Exchange Between Systems—Local and Metropolitan Area Networks—Specific Requirements—Part 3: Carrier Sense Multiple Access with Collision Detection (CSMA/CD) Access Method and Physical Layer Specifications*, Institute for Electronics and Electrical Engineers, Piscataway, NJ, 1998.

6

Wireless Local Area Networks (WLAN)

Jim Paviol
RF Consultant

Carl Andren
John Fakatselis
WLAN Consultant

Wireless local area networks (WLANs) use radio transmissions to substitute for the traditional coaxial cables used with wired LANs like Ethernet. The first generation WLAN products were targeted as wired-LAN extensions. They were originally intended to save money on relocation expenses and demonstrated the utility of wireless laptop operations. Wireless data technology and its application to local-area networks introduced mobile computing. Centrally controlled wireless networks are most often used as part of a larger wired network. A radio base station or Access Point (AP) arbitrates access to the remote wireless stations by means of packetized data. In contrast, in a Peer-to-Peer wireless network, an ad hoc network can be formed at will by a group of wireless stations. New forms of network access protocols, such as Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA) are needed for low error rate operation of wireless networks. Roaming is one of the main advantages of wireless networks, allowing users to freely move about while maintaining connectivity.

The WLAN is used in four major market segments, "Vertical" with factory, warehouse, and retail uses; "Enterprise" with corporate infrastructure mobile Internet uses; "SOHO" (Small Office/Home Office) with small rented space businesses; and "Consumer" with emerging uses. General WLAN trends are shown in Figure 6.1.

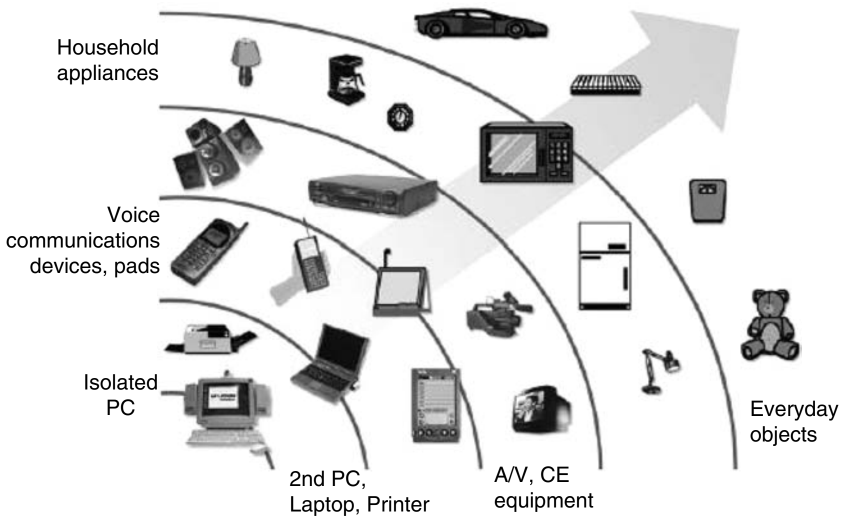


FIGURE 6.1 Commercial WLAN evolution.

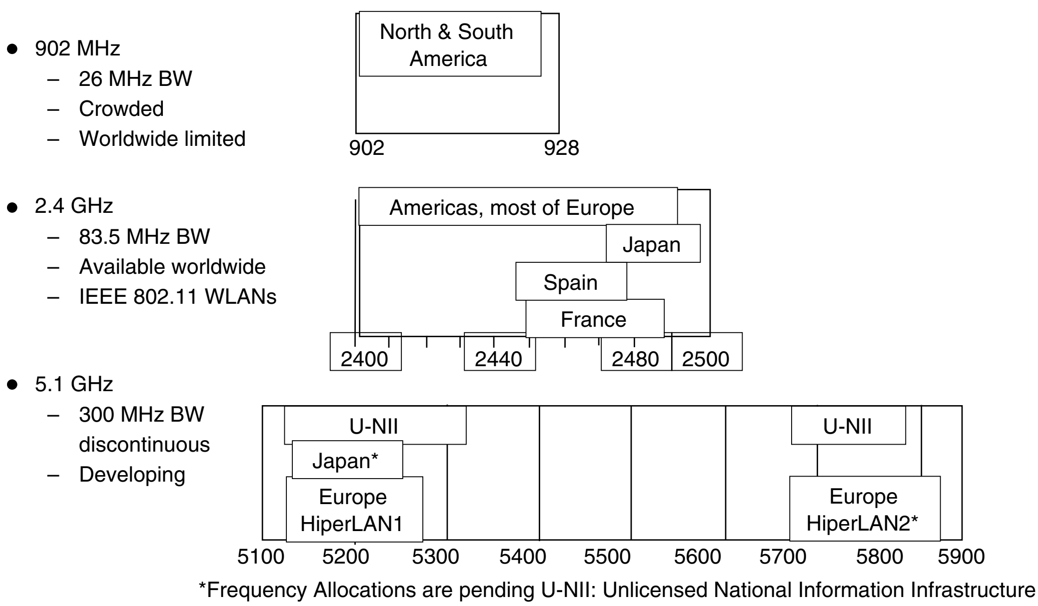


FIGURE 6.2 Unlicensed ISM RF band details.

6.1 WLAN RF ISM Bands

In 1985 the Federal Communications Commission (FCC) in the United States defined the ISM (Industrial, Scientific, and Medical) frequency bands allowing unlicensed spread-spectrum communications. Three of the ISM bands are illustrated in Figure 6.2 with frequencies at 900 MHz, 2.4 GHz, and 5 GHz.

Most important to ISM band WLAN users is that no license for operation is required when the signal transmission is per the guidelines specified by the FCC or other regulatory agencies. Spread-spectrum technology in the ISM band is used to minimize interference and offers a degree of interference immunity

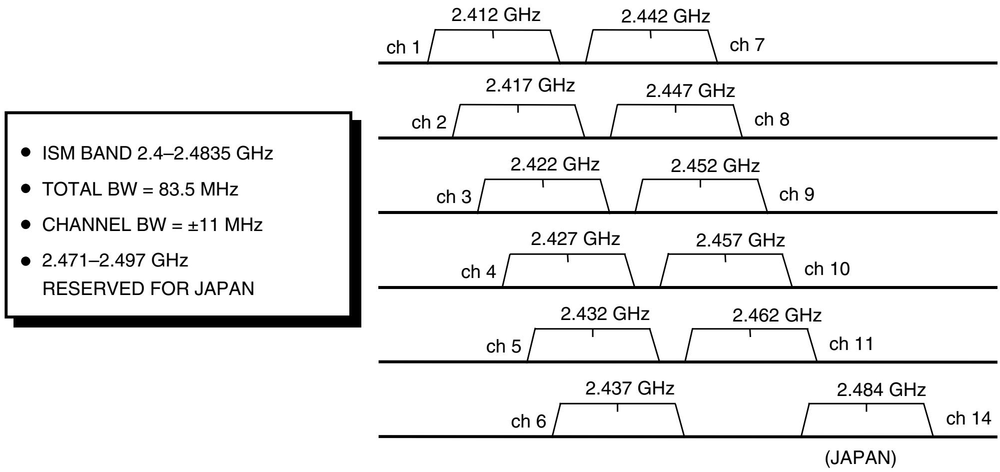


FIGURE 6.3 Operating channels for direct sequence.

from other jamming signals or noise. Other non-spread commercial applications have existed in the ISM bands for many years such as microwave ovens at 2.4 GHz. This is a major potential interference to WLAN in the 2.4 GHz ISM band and has been accounted for in the system design.

The IEEE 802.11 committee selected the 2.4 GHz ISM band for the first WLAN global standard. Unlike the 900 MHz band, 2.4 GHz is available worldwide; 2.4 GHz also has more available bandwidth than the 900 MHz band, and will support higher data rates and multiple adjacent channels in the band. In comparison with the 5.7 GHz band, it offers a good balance of equipment performance and cost. Increasing the transmit frequency impacts the power dissipation, availability of parts/processes and limits the indoor range. The 2.4 GHz band is ideal for a WLAN high-speed, unlicensed data link.

The 900 MHz band has been in use for some time, and component prices are very reasonable. Many cordless phones use this band. The 900 MHz band is quite crowded and it does not have global spectrum allocation. The 2.4 GHz band is less crowded, has global allocations, and the associated technology is very cost effective. This is the band in which IEEE 802.11b, Bluetooth, and HomeRF operate.

The 2.4 GHz band is most heavily used for WLAN. Operating channels are shown in Figure 6.3.

The 5 GHz band has two 100 MHz segments for unlicensed use collectively known as the Unlicensed National Information Infrastructure (UNII) bands. There is a similar allocation in Europe, but it is currently reserved for devices that operate in compliance with the HIPERLAN standard. 5 GHz components are more expensive, and radio propagation has higher losses and more severe multipath at these frequencies. These impairments can be overcome, and systems offering data rates in excess of 50 Mbps in the 5 GHz band will become mainstream in the next several years.

6.2 WLAN Standardization at 2.4 GHz: IEEE 802.11b

Wired LAN standards were developed by the IEEE 802 committee such as IEEE 802.1 Systems Management/Networking, IEEE 802.3 Ethernet, IEEE 802.4 Token Ring, and IEEE 802.6 Metropolitan Area Networks. In 1990, the IEEE 802.11 Wireless LAN Working Group was formed and has in excess of 100 active voting members with global representation. It ratified the 802.11b high rate 2.4 GHz WLAN standard in 1999. The 802.11 standard fostered development of interoperable, inexpensive, and flexible equipment in the 2.4 GHz ISM band. Specified data rates for the IEEE 802.11 2.4 GHz WLAN are 1, 2, 5.5, and 11 Mbps. Spread-spectrum technology is specified in the 802.11 standard transceiver to provide a robust solution in a multi-user environment. One advantage to spread-spectrum techniques in the ISM bands are seen in the allowable transmit power levels. System transmitter power for IEEE 802.11 WLANs must conform to a regulatory agency's specified levels for unlicensed operation. As an example, the FCC states that non-spread-spectrum applications in this band are limited to a 50 mV/m at 3 m. This translates into 0.7 mW into a dipole antenna. Spread-spectrum applications in the United States are allowed up to 1 watt of transmit power, clearly giving it a higher signal strength advantage over non-spread systems. The low spectral power density of a spread-spectrum system also limits interference to other in-band users.

Segmentation of a data communication system into layers allows different approaches from various vendors as long as the responsibilities of the individual layer are met. The IEEE 802.11 specification focuses on the Media Access Control (MAC) and Physical (PHY) layers for WLANs.

The MAC layer controls the protocol and physical layer management. The protocol used for IEEE 802.11 WLANs is the CSMA/CA (Carrier Sense Multiple Access/Collision Avoidance).

The Physical Layer controls the wireless transmission and reception of digital data from the MAC. It is the transceiver or radio for the WLAN. IEEE 802.11 specifies three different physical layer options, Direct Sequence Spread Spectrum (DSSS), Frequency Hopping Spread Spectrum (FHSS), and Diffused Infrared (DFIR). The DFIR method has the shortest range of operation and is limited to indoor operation due to interference from sunlight. DSSS and FHSS are RF technologies that must conform to the standards set by the regulatory agencies of various countries such as the FCC. These impact items such as the allowable bandwidth and transmit power levels.

Another feature of IEEE 802.11 is that the data is packetized. Packetized data is a fixed number of data bytes sent in a single radio transmission of finite length. The data is grouped in frames up to 2304 bytes in length. A common data length is 1500 bytes. A header and preamble are attached in front of the data frame for control information. The preamble is the initial sequence at the start of the radio transmission that allows the demodulator to synchronize its timing and recognize key information concerning the data that follows. Short and long preambles exist. These are specified in greater detail within the standard. The packetization supports the CSMA/CA protocol.

6.3 Frequency Hopped (FH) versus Direct Sequence Spread Spectrum (DSSS)

Frequency hopped (FH) uses a form of FSK modulation called GFSK (Gaussian Frequency Shift Keying). The baseline 1 Mbps data rate for FH IEEE 802.11 has a 2-level GFSK modulation scheme. The symbol {1} is the center carrier frequency plus a peak deviation of (f+), whereas the symbol {o} is the center carrier frequency minus a peak deviation of (f–). The carrier frequency hops every 400 ms over the channel bandwidth per a prescribed periodic PN code. This channel is divided into 79 sub-bands. Each sub-band has a 1 MHz bandwidth. The minimum hop rate of 2.5 hops/s allows several complete data packets or frames to be sent at one carrier frequency before a hop.

Direct sequence spread spectrum (DSSS) in IEEE 802.11 specifies a DBPSK and DQPSK (D = Differential) modulation for 1 and 2 Mbps data rates. Differential techniques use the received signal itself to demodulate the signal by delaying one symbol period to obtain clock information. In DBPSK, a logic 1-bit input initiates a 180-degree phase change in the carrier and a ϕ-bit initiates no phase change in the carrier. DQPSK has a 0-, 90-, or 180-degree phase transition on each symbol.

The carrier frequency in an IEEE 802.11 DSSS transmitter is spread by an 11-bit Barker code. The chipping rate is 11 MHz for a 1 Mbit data rate. This yields a processing gain of 11. The main lobe spacing is twice the chip rate and each side lobe is the chip rate as shown in Figure 6.4. The DSSS receiver will filter the side lobes, downconvert the main lobe spectral component to baseband, and use a copy of the PN code in a correlator circuit to recover the transmitted signal. The FH scheme has been limited to 1 Mbit and 2 Mbit by technical and regulatory issues. This is expected to limit the use of FH systems compared with the more versatile high-rate DSSS modulation in future applications. An illustration contrasting frequency hopping to spread spectrum is shown in Figure 6.4.

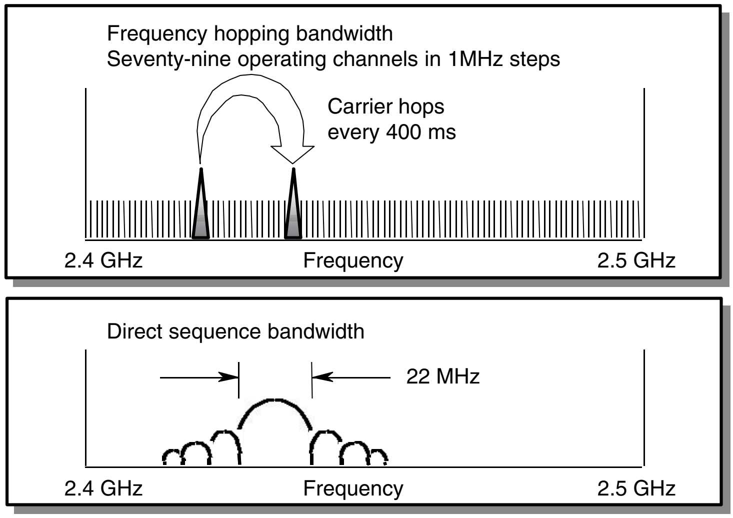


FIGURE 6.4 IEEE 802.11 frequency hopping versus direct sequence spread spectrum.

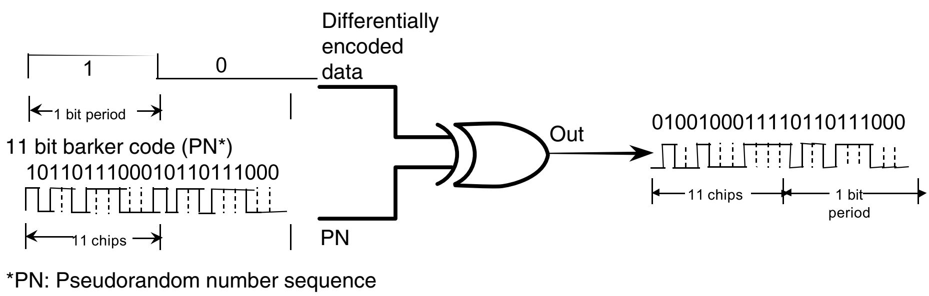


FIGURE 6.5 Direct sequence spectrum spreading. RF energy is spread by XOR of data with PN sequence.

6.4 Direct Sequence Spread Spectrum Energy Spreading

In DSSS systems the spreading of the data is achieved by modulating the data with a Pseudo-random Number (PN) sequence of binary values called a PN code. If the PN code has a bandwidth much higher than the bandwidth of the data (approximately ×10 or greater), then the bandwidth of the modulated signal will have a spectrum that is nearly the same as a wideband PN signal. An 11 bit Barker Code is used with the IEEE 802.11 DSSS PN spreading. A Barker code was chosen for its unique short code properties.

By multiplying the information-bearing signal by the PN signal, each information bit is chopped up into a number of small time increments called "chips," as illustrated in the waveform diagram shown in Figure 6.5. The rate at which the PN code is clocked to spread the data is called the chip rate. The PN sequence is a periodic binary sequence with a noise-like waveform. The acquisition of the data in the receiver is achieved by correlation of the received signal with the same PN code that was used to spread the signal at the transmitter. In DSSS systems the data is primarily PSK modulated before spreading. The spreading produces a "Processing Gain" dependent upon the PN spreading code to symbol rate ratio. This value is a minimum of 10 dB for the IEEE 802.11 DSSS WLAN waveform. The spectrum after this PN code is used is wider and lower in signal level as illustrated in Figure 6.6.

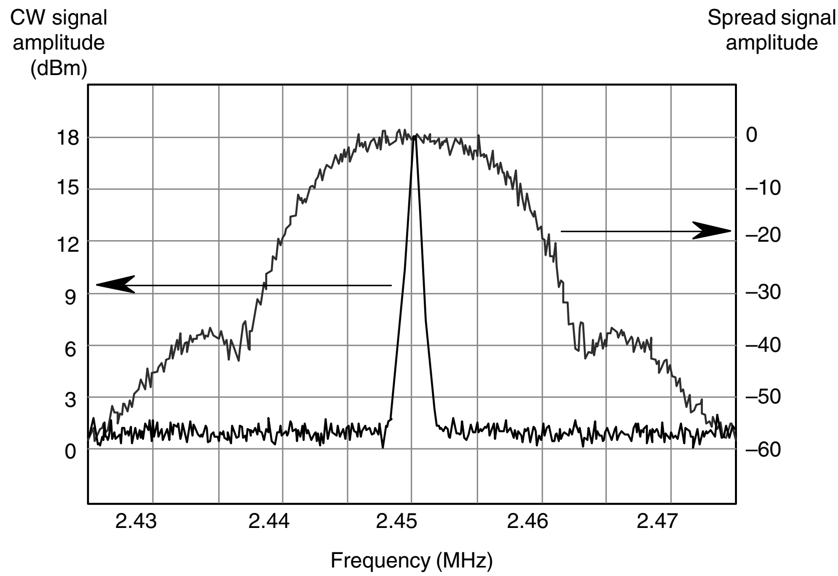


FIGURE 6.6 Direct sequence spread spectrum.

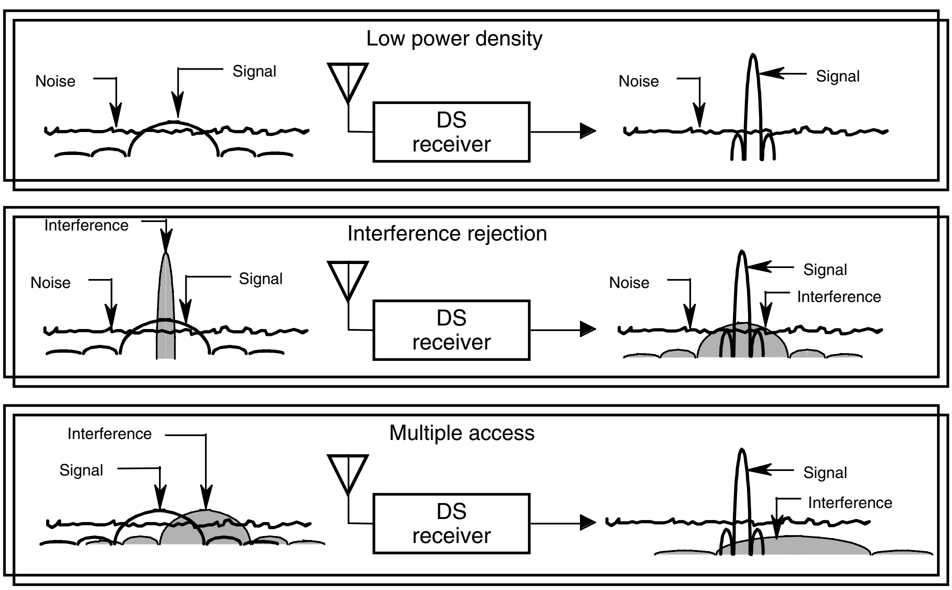


FIGURE 6.7 Direct sequence spread spectrum properties.

The primary advantage of a spread-spectrum system is its ability to reject interference whether it be the unintentional interference by another user transmitting on the same channel, or the intentional interference by a hostile transmitter attempting to jam the transmission. Due to its spread characteristic, a DSSS signal appears as noise to all receivers except the one meant to receive the signal. The intended receiver is able to recover the spread signal by means of correlation, which simultaneously recovers the signal of interest and suppresses the background noise by the amount of processing gain.

The reception of the DSSS signal in the presence of a narrow band interferer is accomplished by de-spreading the signal of interest while spreading the energy of the interfering signal by the amount equal to the processing gain. The result of the de-spreading process with an interference source is illustrated in Figure 6.7 for narrow band interfering signals. Likewise, recovery of the signal of interest in presence of another spread signal having a different PN code is achieved by further spreading the interference while de-spreading the signal with the matching PN code. The worst case interference to DSSS systems is a narrow band interference signal in the middle of the spectrum of the spread signal.

6.5 Modulation Techniques and Data Rates

The 1 Mbit data rate is formed using Binary Phase Shift Keying (BPSK) and the 2 Mbit data rate uses Quadrature Phase Shift Keying (QPSK). QPSK doubles the data rate by increasing the number of bits per symbol from one (BPSK) to two (QPSK) within the same bandwidth. Both rates use BPSK for the acquisition header known as the preamble.

The high rate modulation for 5.5 and 11 mbps uses a form of M-Ary Orthogonal keying called Complementary Code Keying (CCK). CCK provides coding gain by the use of modified Walsh functions applied to the data stream. Two forms of this are used to provide multiple rates for stressed links. Altogether, four data rates are provided. To make 11 Mbit CCK modulation, the input is formed into bytes and then subgrouped into 2 and 6 bits. The 6 bits are used to pick one of 64 complex vectors of 8-chip length and the other 2-bits DQPSK modulate the whole symbol vector. For 5.5 Mbit CCK mode, the incoming data is grouped into 4 bit nibbles where 2 of those bits select the spreading function out of a set of 4 (the 4 having the greatest distance of the 11 Mbit set) while the remaining 2 bits set the QPSK polarity of the symbol. The spreading sequence modulates the carrier by driving the I and Q modulators. Figure 6.8 illustrates modulation at the four data rates of 1, 2, 5.5, and 11 Mbit.

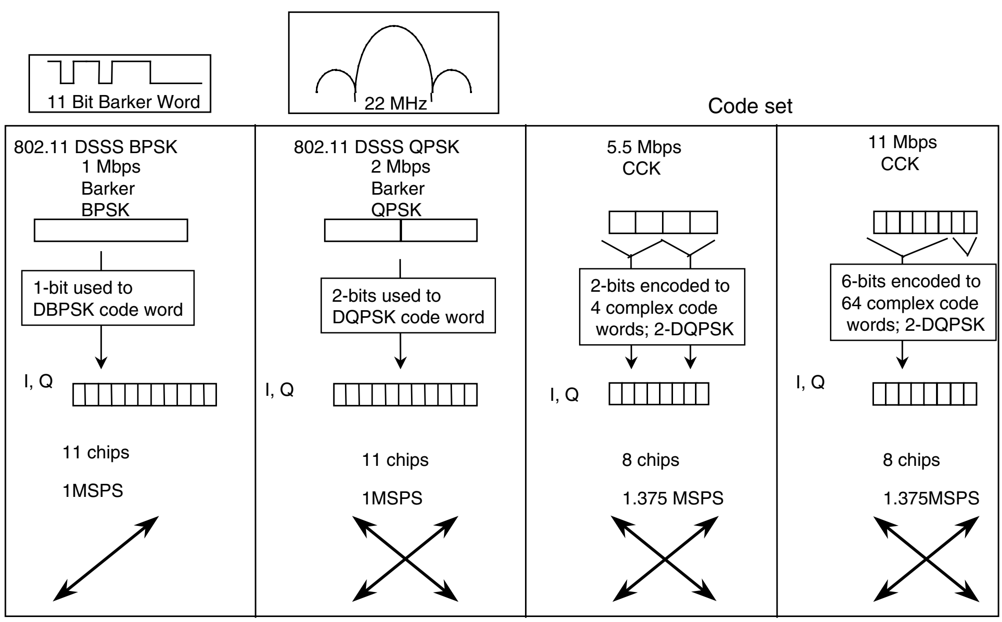


FIGURE 6.8 Modulation techniques and data rates.

6.6 Carrier Sense Multiple Access/Collision Avoidance

The wireless environment offers a greater challenge to the WLAN designer when compared with the wired LAN environment. Wired LANs use a CSMA/CD (Carrier Sense Multiple Access/Collision Detect) protocol. Data collisions in a wired environment produce a unique voltage that can be monitored. This allows for random back-off time periods before the system initiates a data resend.

IEEE 802.11 WLANs use the CSMA/CA (Collision Avoidance) protocol. This protocol avoids data collisions by having the system listen to the channel and wait before sending a message. With CSMA/CA, only one node may talk at a time. This highlights the importance of performing a Clear Channel Assessment (CCA) to determine if the medium or channel is clear to transmit. Although the MAC controls the CSMA/CA protocol, the responsibility falls upon the physical layer to perform CCA.

IEEE 802.11 states that the physical layer must be able to provide at least one of three specified methods for CCA. CCA mode 1 simply detects energy above a programmable level. If no signal is present, the channel is clear to transmit. If the signal is present, the system will wait a set time period to check the channel again. CCA mode 2 provides the carrier sense function. Since this is a spread spectrum system, correlating the received signal with the 11-bit PN code performs carrier sense. No correlation indicates that the channel is clear to transmit. Correlation with a signal shows that the channel is busy and that the system will back off for a time period and try again. CCA mode 3 combines modes 1 and 2 by reporting a busy medium with both detection of energy and carrier sense. Figure 6.9 illustrates a four-station scenario where radio collisions are avoided using CSMA-CA.

6.7 Packet Data Frames in DSSS

The 802.11 WLAN standard specifies data packetization. This means that the data is segmented into frames with a preamble and header attached at the start of each frame. The preamble allows carrier and correlation lock as well as user identification. The header contains management and control information for the data transmission.

The sync field is made up of 128 one bits. Note that all bits are processed by a scrambler function, which is part of the IEEE 802.11 spread spectrum physical layer so this original pattern of ones will be altered. The purpose of the sync field is to allow the receiver to lock on to the signal and to correlate to the PN code.

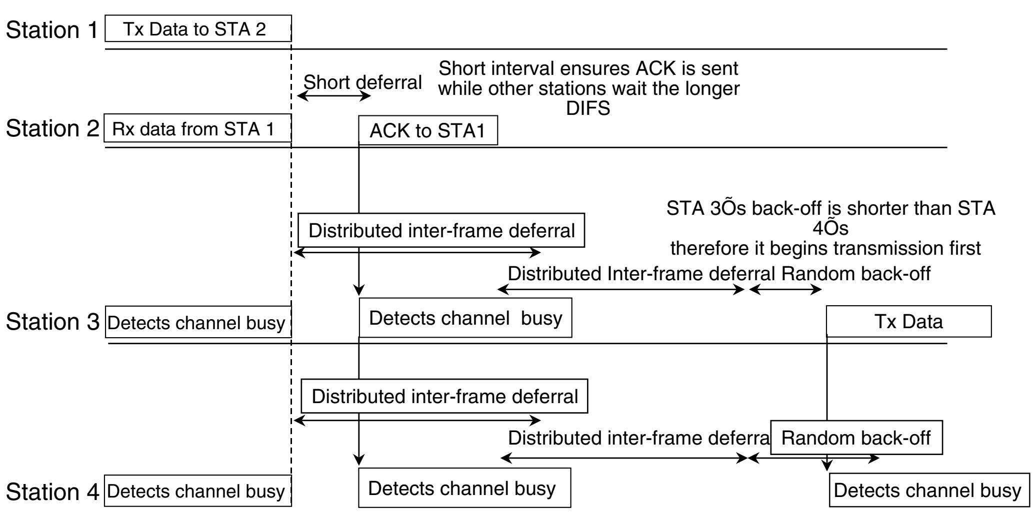


FIGURE 6.9 Carrier sense multiple access/collision avoidance.

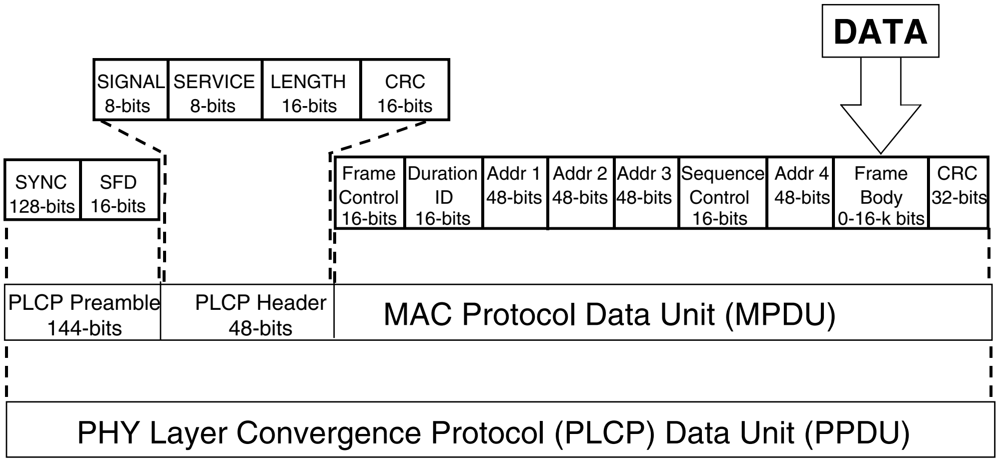


FIGURE 6.10 Frame format.

The start frame delimiter (SFD) field initiates the start of the data frame. The signal field indicates the data packet data wave. Note that the preamble and header are always transmitted as a DBPSK waveform. The length field defines the data packet size with a maximum length of 2304 bytes. Finally, one CRC (Cyclic Redundancy Check) protects the signal, service, and length fields with a frame check sequence and another protects the payload.

At the end of the data packet, the receiver will send an acknowledgment (ACK) indicating successfully transmitted data. If a data packet were lost either by multipath fading or interference, the sender would retransmit the packet. Figure 6.10 details frame format, preamble, header, and data.

6.8 IEEE 802.11 Network Modes

There are two network modes: Ad Hoc and Infrastructure. Ad Hoc mode permits users within range to set up a network among them without any infrastructure. Infrastructure mode uses an access point (AP) to coordinate the users and allow access to the wired network services.

6.8.1 Ad Hoc Mode

The distributed coordination function (DCF) forms the basis to implement Ad Hoc networking. The provisions in the standard allow the creation and dissolution of an Ad Hoc network to be straightforward for users to set up. The CSMA/CA medium access method provides for fair access to the radio channel among all of the users. All data exchanges are directly between the individual stations.

When a single station is designated as the coordination function, the network is known as a basic service set (BSS). This is illustrated in the peer-to-peer AdHoc network in Figure 6.11.

6.8.2 Infrastructure Mode

An AP is a device that connects the wireless stations to the distribution system in a network. It typically will be configured to be the single Coordination Function in a BSS. The network planning for a large installation involves site surveys to do the cellular radio planning needed to determine the number and location of APs. The channel assignments for each AP can be optimized to reduce interference from adjacent cells depending on the physical layout. Many times, the installation will utilize a wired Ethernet network to connect a number of APs to the network server. Each AP will manage the traffic within its BSS. Stations in adjacent cells will recognize when the packet is not intended for its BSS. The same DCF

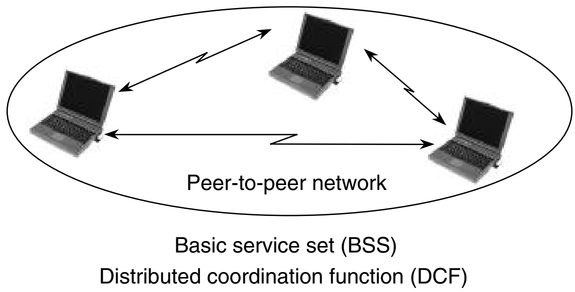


FIGURE 6.11 IEEE 802.11 ad hoc mode.

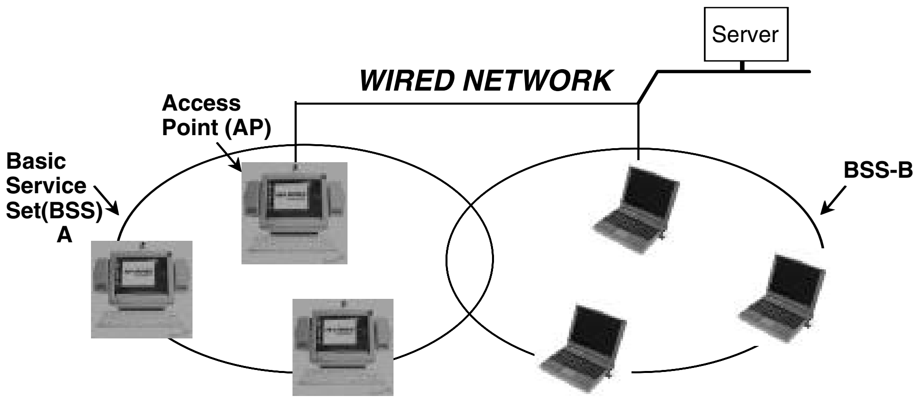


FIGURE 6.12 IEEE 802.11 infrastructure mode.

mechanisms provided in the ad hoc mode help to manage radio interference from adjacent cells. Mechanisms such as Authentication, Association, and Wired Equivalency Privacy (WEP) provide security for the overall network. The authentication process is used to verify the identity of stations that are allowed access to the network. As users move between the various radios cells and they can no longer communicate with their AP, they can scan the channels to look for a new AP to associate with. With proper radio cell-planning, users can be assured of constant coverage as they roam through the facility. An infrastructure mode illustration is shown in Figure 6.12.

6.8.3 "Hidden" Nodes

The planning for a WLAN ranges from none for an ad hoc IBSS, to careful radio surveys for AP-based infrastructure systems in large enterprise computing environments. It is impossible to plan for perfect radio coverage given the uncertain and time-varying conditions in the channel. Movement and location as well as indoor topology will affect radio coverage. In providing complete coverage of the facility there will be situations where there are two stations that can communicate with their common AP, but are out of range to hear each other. This is known as the Hidden Node problem. This could cause additional collisions to occur at the AP, since when one station is transmitting, the other will determine the channel is clear because it is out of range. After the normal deferral time it will transmit and will interfere with a packet from the other station. The standard provides an optional mechanism called Virtual Carrier Sense (VCS) to reduce the collision due to the hidden node problem. This mode is based on using a Request To Send (RTS) and Clear To Send (CTS) exchange between the station and the AP. When the

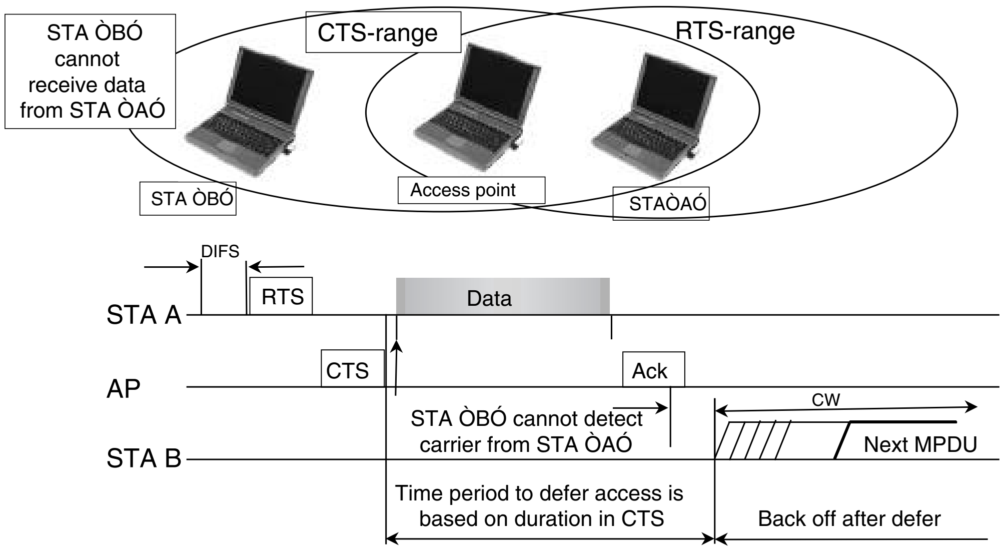


FIGURE 6.13 "Hidden node" provisions.

station has determined it is clear to transmit, it will send a short request to send a message with a transaction duration included. The AP will respond with a CTS that also includes the transaction duration. The other station that is hidden cannot hear the RTS, but can hear the CTS. It will read the duration field and not begin looking for a clear channel until that time has elapsed. Hidden nodes are illustrated in Figure 6.13 showing Stations A and B with an access point.

6.8.4 Point Coordination Function

The point coordination function (PCF) is an optional mode that can be selected in the installation of an 802.11 Network. This mode is provided to optimize the network throughput. Rather than having each station contend for the channel using the CCA and the random back-off periods, the AP defines contention-free periods using a beacon frame sent at a regular interval. The network allocation vector (NAV) is a variable that is transmitted in the control frames to tell all of the stations the duration to defer from accessing the channel. It is used to define the length of the contention-free period. The AP will then poll each station during the contention-free period. The poll will send data if there is some waiting for transmission to that station and request data from the station. As each station is polled, it will acknowledge reception and will include data if there is some pending transmission to the AP. One of the benefits of the PCF mode is that it allows the network planner to improve the probability of the delivery of data in a certain time bound. This is critical for real time data such as voice, audio, video, and multimedia. Minimizing the uncertainties of when a station can get the channel that exists in the DCF mode optimizes the system performance. If a packet of audio or video data has to be retransmitted due to lost packets and extended deferral times, the quality of the audio or video will suffer. The network can be optimized to permit a certain amount of retries within the contention-free period to improve the quality of service to the end application. The PCF timing is illustrated in Figure 6.14.

6.8.5 WLAN Security

Since they can be received outside of the controlled facility, wireless networks are more vulnerable to interference and theft than wired networks. The data security field known as cryptography is rapidly

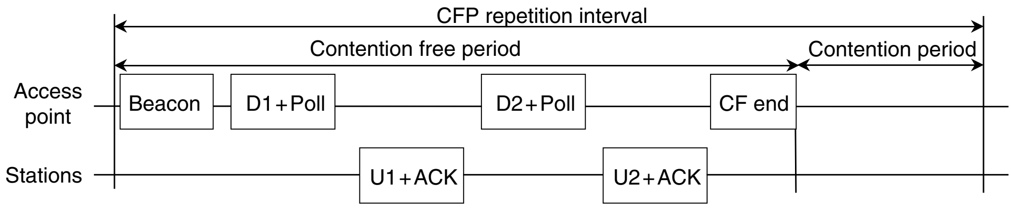


FIGURE 6.14 IEEE 802.11 point coordination function.

growing as more systems convert to wireless operation. Spread spectrum offers a little security by spreading the signal over a wide bandwidth.

There are a number of mechanisms provided in 802.11 to minimize the chances of someone either logging on the network without authorization or receiving and using data received off the air from authorized users. As an option the data can be encrypted to prevent someone who is receiving packets from the network to be able to interpret the data. The keys for the security are distributed to the users in the network by a secure key management procedure. Without the key the snooper will have to resort to complex code-breaking techniques to retrieve the original data. The level of encryption is defined as Wired Equivalency Privacy (WEP). It is strong enough to require effort equivalent to that required to get data from a wired LAN. This algorithm is licensed from RSA Data Security. The encryption mechanism is used in the authentication process as well.

6.8.6 Data Encryption

The WLAN Security: Authentication diagram illustrates a technique for authenticating the identity using the encryption features. In the "Challenge and Response Protocol" shown, the station transmits a "challenge" random message(r) to the AP. The AP receives the message, encrypts it by using a network algorithm (fK1), and transmits the encrypted information (y) "response" (fK1) back to A. System A has access to the network algorithm (Y) and compares it to the received response. If $y = y'$, the identification procedure criteria has been met and data transmission will follow. It does not, then A will issue a new challenge with a different random message to B. Note that A and B share a common (private) key, k1. After the authentication process is successfully completed the station will then associate itself with an AP. The association tells the overall network which AP services any station. Once identification has occurred, both systems communicate using network encryption algorithms. In this case, the station encrypts data (x) and transmits a cipher (y) over the channel. The AP or another station has access to the encryption method and decodes the cipher to obtain the data. Because the stations in a BSS share a common key (k1), this type of encryption is called private-key cryptography.

Also included in the standard is an Integrity Check Vector (ICV). This is a variable added to an encrypted data packet as an additional error-detection mechanism. After the decryption process the transmitted ICV is compared with the ICV calculated from the plain text as an error check. Figure 6.15 illustrates data encryption with a block diagram.

6.9 5 GHz WLAN

With the FCC rule and order establishing the Unlicensed National Information Infrastructure (U-NII), 300 more MHz of bandwidth were made available for WLAN users. The purpose was to expand the access of people to information without a lot of infrastructure build out. One of the frequently sited

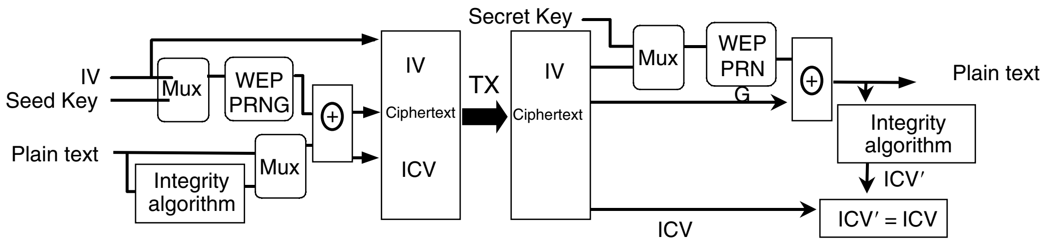


FIGURE 6.15 WLAN security: data encryption.

scenarios is the distribution of Internet access to schools without having to wire classrooms. The FCC chose to put in a minimum of specifications for the waveforms to be used in the band.

Spread Spectrum is not a requirement. Transmit power and power spectral density are the primary specifications. No channelization or spectrum-sharing rules were included in the 15-part regulation. It remains to be seen how the various devices and standards will coexist in this new band. The band is split into three 100 MHz bands for defining maximum output power and spurious emissions levels. There are users of licensed bands on all sides of these new bands that argued for protection against interference from unlicensed devices. This drives the channel band edges for carriers and defines radio requirements for suppression of spurious emissions. The power levels allowed are 50 mW, 250 mW, and 1 W in the lowest, middle, and upper bands, respectively. A change was made in 1998 to permit higher antenna gain with a reduction in output power. This extends the range for point-to-point links.

In Europe 150 MHz of bandwidth is set aside for HIPERLAN1 devices. These are WLAN devices with the same functional requirements as an 802.11a WLAN. As opposed to the FCC, the ETSI regulation requires all devices to meet the HIPERLAN standard for the PHY and MAC layers. The maximum data rate is 54 Mbit and the modulation is GMSK. Equalization is required for reliable operation. A small number of HIPERLAN products have been introduced since the standard was approved.

ETSI is defining a standard for the 5 GHz band that is oriented to wireless ATM traffic. This is suitable for Quality of Service applications and for wireless connectivity to an ATM system. The final spectral allocation has not been made for these devices.

Japan has started the development of standards for WLAN devices in the 5 GHz band. They have decided that devices will be required to use the same physical layer implementation that is defined in the IEEE 802.11a standard.

6.10 RF Link Considerations

The radio link performance can be characterized as consisting of radio design variables and link variables. When all the variables are understood the link performance can be determined. The most important WLAN link performance measure is the packet error rate (PER), which is usually expressed as a percentage. Most radio link parameters are given at a packet error rate of 0.1 or 10 percent. This is the highest practical acceptable packet error rate and provides the design maximum. The IEEE 802.11 standard specifies an 8% PER for measurements.

All of the radio design variables combine to provide a given performance on a given link. The transmit power together with the receiver noise floor determines the ultimate range of the radio. Higher transmit power provides greater range but also higher battery drain. The antennas for wireless systems are generally omni-directional to allow mobility. The higher data rate requires a higher signal-to-noise ratio for the same error rate. The bit error rate and the length of the packet determine the packet error rate. Longer packets require lower bit error rates. Different modulation schemes require more or less power to achieve the same bit error rate. For instance FSK requires more power to achieve the same bit error rate as CCK. The link variables include the range, which is the distance from the transmitter to the receiver, the multipath environment, and the interference environment. Missed packets and corrupted (including recoverable) packet data are included in PER.

Radio signals radiated by an ideal isotropic antenna weaken with the square of the distance as they travel through free space (square power law). The attenuation also increases with frequency. At 2.4 GHz (Lambda = 0.125 m = 5”) the path loss in free space is about 40dB for 1m. Propagation of RF signals in the presence of obstacles are governed by three phenomena: reflection, diffraction, and scattering. Reflection occurs when the dimension of obstacles are large compared to the wavelength of the radio wave. Diffraction occurs when obstacles are impenetrable by the radio wave. Based on Huygen’s principle, secondary waves are formed behind the obstructing body even though there is no line of sight. Scattering occurs where the obstacles have dimensions that are on the order of the wavelength. The three propagation mechanisms all have impact on the instantaneous received signal in all different directions from the transmitting antenna.

Measurements have shown that propagation loss between floors does not increase linearly (in dB) with increasing separation of floors. Rather, the propagation loss between floors starts to diminish with increasing separation of floors. This phenomenon is thought to be caused by diffraction of radio waves along the side of a building as the radio waves penetrate the building’s windows. Values for wall and door attenuation are shown in Table 6.1 and a plot of attenuation between building floors is shown in Figure 6.16.

TABLE 6.1 Indoor Propagation Path Loss

2.4 GHz signal attenuation through:	
Window in brick wall	2 dB
Metal frame, glass wall into building	6 dB
Office wall	6 dB
Metal door in office wall	6 dB
Cinder wall	4 dB
Metal door in brick wall	12.4 dB
Brick wall next to metal door	3 dB

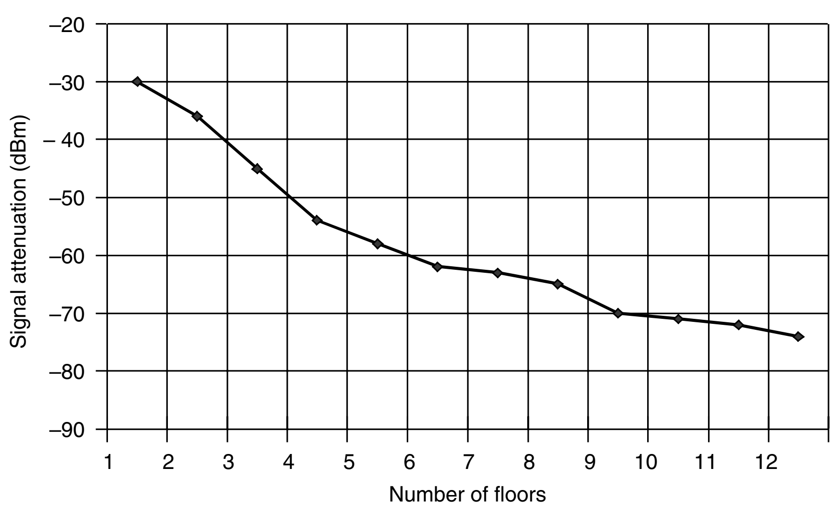


FIGURE 6.16 Path loss between building floors.

6.10.1 WLAN Power, Sensitivity, and Range

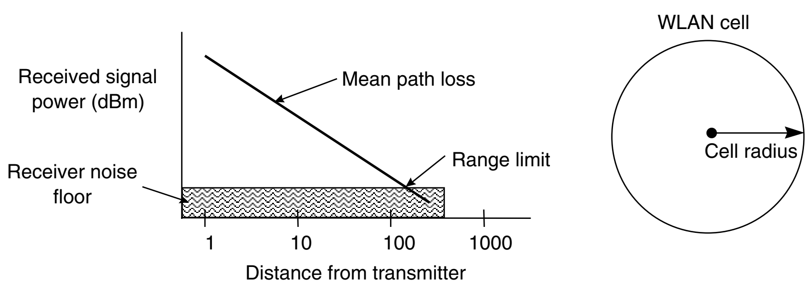
The range of a WLAN radio is influenced by data rate, bit error rate requirements, modulating waveform S/N ratio, power amplification, receiver sensitivity, and antenna gain. At higher data rates more power is required. The log of the data increase is the amount of power increase in dB to achieve the same range. Reliability is also a function of power. Higher power produces a lower bit error at the same range, or more range at the same bit error rate. The waveform is a significant contributor to performance. Phase Shift Keying (PSK) type waveforms are the most power efficient. Frequency Shift Keying (FSK) requires almost twice as much transmitted power to achieve the same range.

IEEE 802.11 sensitivity for 11 Mbit CCK QPSK is specified as a minimum of –76 dBm. Typical radios attain 6 dB better than the minimum levels. Lower data rates such as 5.5 Mbit CCK and 2 Mbit QPSK have better sensitivities by approximately 5 dB than the 11 Mbit levels. The best sensitivity is obtained with the 1 Mbit BPSK. Values are typically 3 dB better than the 2 Mbit, QPSK thus 1 Mbit is used for the header/preamble information.

In an ideal propagation environment such as free space (i.e., no reflectors), the transmitted power reaches the receiver some distance away attenuated as a function of the distance, r. As shown in Figure 6.17, the Constant, 40.2, is used for 2.4 GHz propagation and changes insignificantly across the ISM band. The exponent of 2 is for free space and increases with multipath. As the receiver moves away from the transmitter, the received signal power reduces until it dips into the receiver noise floor, at which time the error rate becomes unacceptable. This is the first order determination of the largest a cell can be. This model can be used for unobstructed line-of-sight propagation with highly directional antennas where the antenna gain allows a propagation path that is miles long.

6.10.2 Signal Fading and Multipath

As a transmitted radio wave undergoes reflection, diffraction, and scattering it reaches the receiving antenna via more than one path giving rise to a phenomenon called multipath. The multiple paths of the received signals cause them to have varying signal strengths as well as having different time delays (phase shifts), also known as delay spread. These signals are summed together (vector addition) by the receiving antenna according to their random instantaneous phase and strength giving rise to what is known as small-scale fading. Small-scale fading is a spatial phenomenon that manifests itself in the time domain having Rayleigh distribution; hence it is called Rayleigh fading. Small-scale fading produces



Loss dB = 40.2 + 10 · log (r^n)

- Where Loss is transmit power/received power in dB
  – r is the cell radius
  – n is 2 for free space

FIGURE 6.17 Path loss — free space. Loss dB = 40.2 + 10 • log (r^n) where Loss is transmit power/received power in dB; r is the cell radius; and n is 2 for free space.

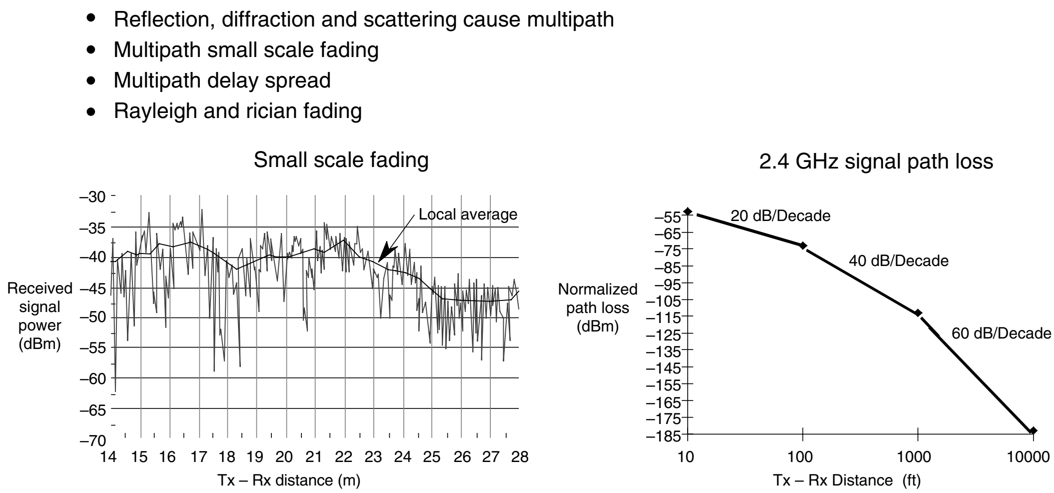


FIGURE 6.18 Signal fading and multipath. Reflection, diffraction, and scattering cause multipath; multipath small scale fading; multipath delay spread; Rayleigh and Rician fading.

instantaneous power levels that may vary as much as 30 or 40dB while the local average signal level changes much more slowly with distance.

Just as the power law relationship between distance and received power is applied to path loss in free space, it may be used in the presence of obstacles. A general propagation loss model for local average received power uses a parameter, n, to denote the power law relationship where n = 2 for free space, and is generally higher for indoor wireless channels.

The "2.4 GHz Signal Path Loss" curve in Figure 6.18 represents various path losses with different values of n. The first segment, n = 2, of the curve, loss is primarily free space loss. The second and last segments of the curve have values of 4 and 6 for n, respectively, representing more lossy channels. The instantaneous drop of the signal power as it transitions from –40 dB/dec to –60 dB/dec is typical of a signal loss when a receiver loses line-of-sight to its respective transmitter.

In a multipath condition where the receiver also has a line-of-sight path to the transmitter, the statistical distribution of the local average signal level follows Rician distribution. Rician distribution is based on a factor, k, which specifies the ratio of direct path versus multipath power levels. Multipath is illustrated in Figure 6.18.

6.10.2.1 Log Normal and Rayleigh Fading

The mechanism of the multipath fading can be viewed as being caused by two separate factors: the product of the reflection coefficients and the summation of the signal paths. These two mechanisms produce separate fading characteristics and can be described by their probability distribution functions. The first is characterized as having a Log Normal distribution and is called Log Normal fading. The second mechanism, the sum of the signal paths, produces a Rayleigh probability distribution function and is called Rayleigh fading. Figure 6.19 illustrates both multipath mechanisms.

Significant effort has gone into characterizing the multipath environment so that effective radio structures can be designed that operate in difficult high reflection environments.

6.10.2.2 Effects of Multipath Fading

The value of (n) is 2 for free space propagation, but in general, (n) can take non-integer values greater or less than 2 depending on the environment. k is a log normal random variable that is added to model the variability in the environment due to the different amounts and types of material through which the signal travels. The uncertainty is shown in Figure 6.20.

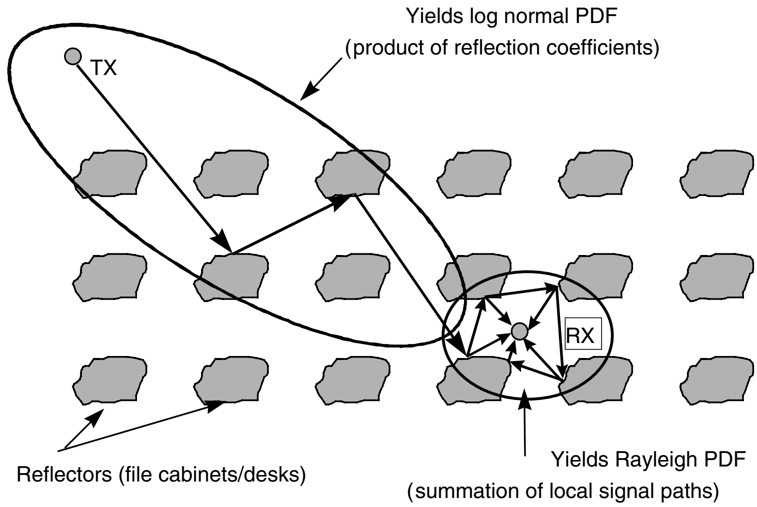


FIGURE 6.19 Log normal and Rayleigh fading.

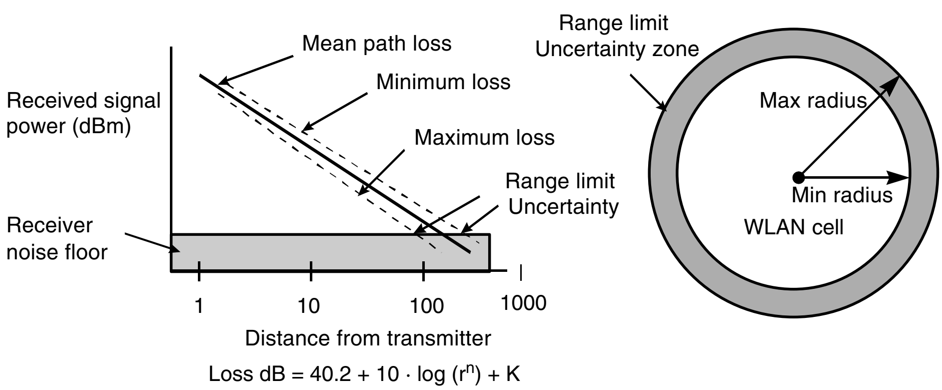


FIGURE 6.20 Effects of multipath fading.

Residential: n = 1.4–4.0, with n = 2.8 typical
Residential: Standard deviation of the log normal distribution 7–12 dB with 8 dB typical
Office: n = 1.74–6.5, with n = 3.7 typical
Office: Standard deviation of the log normal distribution 6–16 dB with 10 dB typical
Light industrial: n = 1.4–4.0, with n = 2.2 typical (open plan),
Light Industrial: Standard deviation of the log normal distribution: 4–12 dB with 10 dB typical

6.10.2.3 Delay Spread Craters

Multipath fading has been the chief performance criteria for selecting a new high-data-rate 802.11 DSSS standard waveform. Many independent multipath surveys published in the IEEE literature are in agreement in showing that high delay-spread holes exist anywhere within a cell. These holes can be a real difficulty for cell planners because stations may fail to operate even a short distance from the cell center. In commercial environments it is common to see the multipath spread reach 100 nsec (rms). Multipath spread is one unit of measurement for the Rayleigh fading characteristic. A 100-nsec multipath spread is commonly observed in cafeterias, atriums, and open Wal-mart-like structures. Craters are illustrated in Figure 6.21.

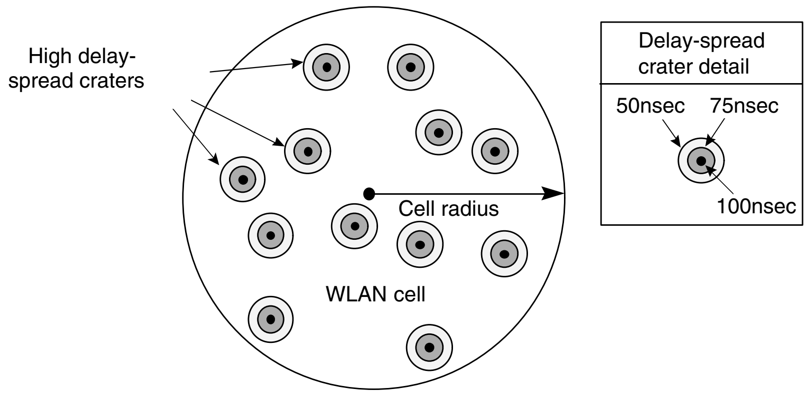


FIGURE 6.21 Delay spread craters.

6.10.2.4 Multipath Mitigation

By itself, antenna diversity provides minimal relief from multipath craters. If a station is located within a crater, two antennas tend to see the same degree of multipath.

In a multipath crater, the SNR is often high. The average signal power in the crater is the same as described by the mean path loss shown before. If the crater is not near the cell boundary, the SNR is good. However, the multipath components can highly distort the signal so the conventional receiver still fails.

A simple parallel is the distortion caused by audio echoes. A speaker becomes unintelligible when severe echoes exist, even in the absence of other noise. The equivalent effect occurs in a receiver. The multipath echoes cause the code words and DSSS spreading chips to overlap in time. This self-interference causes receiver paralysis even in the absence of noise, unless the receiver has been designed to correct the echoes.

6.10.2.5 Impulse Response Channel Models

In a room environment, two measurements of the same radio in the same place may not agree. This is due to the changing position of the people in the room and slight changes in the environment which, as we have seen, can produce significant changes in the signal power at the radio receiver. A consistent channel model is required to allow the comparison of different systems and to provide consistent results. Simulations may be run in software against models of the radio. But more valuable are hardware simulators that can be run against the radio itself. These simulators operate on the output of the radio and produce a simulated signal for the receiver from the transmitted signal. To be of value for comparison, a standard model should be used.

The IEEE 802.11 committee has been using quite a good simulation model that can be readily generalized to many different delay spreads.

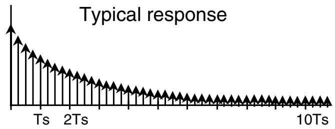
Another model, which is more easily realized in hardware simulations, is the JTC '94 model. This model is a standard that provides three statistically based profiles for residential, office, and commercial environments. Two obscured line-of-sight profiles are provided for residential and commercial environments; these are illustrated in Figure 6.22.

6.10.3 Interference Immunity and Processing Gain

Another significant parameter to be considered in the link is interference. In the ISM band the main source of interference is the microwave oven. Radios must be designed to operate in the presence of microwave ovens. The spread spectrum nature of the waveform allows narrowband interference to be tolerated. Processing gain of 10 dB is available to provide protection from narrowband interferers of any

IEEE 802.11 model

- Ideal for software simulation -> continuously variable delay spread
- Largest number of paths, min 4 per symbol
- Purely exponential decay – no OLOS profiles



JTC Ô94 model

- Ideal for hardware simulation (real time)
- Up to 8 paths
- 9 statistically based profiles, 3 each residential, office, commercial
- 2 OLOS profiles (Residential C and commercial B)

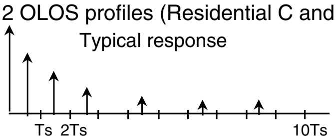


FIGURE 6.22 Impulse response channel models.

type. Rate changes to lower data rates can be used to allow higher tolerance to microwave energy. The IEEE 802.11 protocol is designed to enable operation between microwave energy pulses. Within the 802.11 protocol is the ability to change frequencies to avoid a problem channel.

Other radios in the band can cause interference. Two types of interference must be considered. First is co-channel interference of our own system. This is the energy from a nearby cell of the same system that is on the same frequency. Frequency planning and good layout of access points can minimize this interference and keep it from being a system constraint. The second type of interference is from other systems. These might be Direct Sequence Spread Spectrum or Frequency Hop Spread Spectrum. Two mechanisms are available to mitigate this interference. The first is clear channel assessment (CCA) provided in the IEEE 802.11 standard. MAC layer protocol provides collision avoidance using CSMA-CA. The second is processing gain, which provides some protection from Frequency Hop Spread Spectrum radios, which appear as narrowband interferers. Some frequency planning is always required and not all systems will coexist without some performance degradation.

The radiated energy from a microwave oven can interfere with the wireless transmission of data. The two plots shown in Figure 6.23 illustrate that the radiated energy from a microwave oven is centered in the 2.4 GHz ISM band.

The plots show the results of leakage tests conducted at 9.5 inches and 10 feet from a microwave oven. The test shown on the left plot was conducted at a distance of 9.5 inches from the oven. The leakage test shown on the right plot was conducted at a distance of 10 feet.

The two cases are typical plots selected from more that 20 leakage tests performed on different brands of ovens, at different distances and different antenna angles. The measurements show that the energy drops off dramatically when the distance between the measurement antenna and the oven exceeds six feet. At a distance of ten feet, the energy level is typically below the –20 dBm level as illustrated in the plot on the right.

6.11 WLAN System Example: PRISM® II

The IEEE 802.11b high rate WLAN has been implemented with the Intersil PRISM® (Personal Radio using Industrial Scientific Medical bands) SiGe chipset. The block diagram is shown in Figure 6.24. This

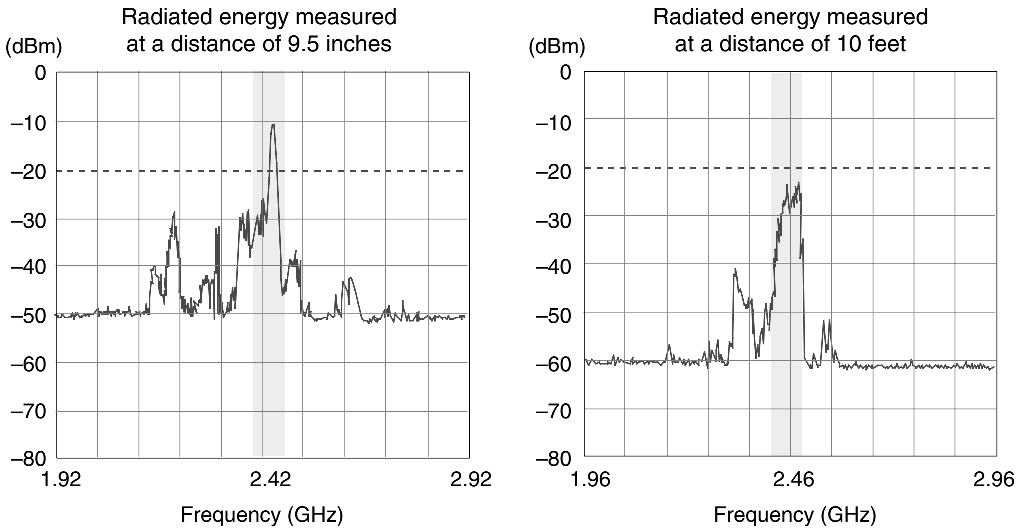


FIGURE 6.23 Microwave oven interference.

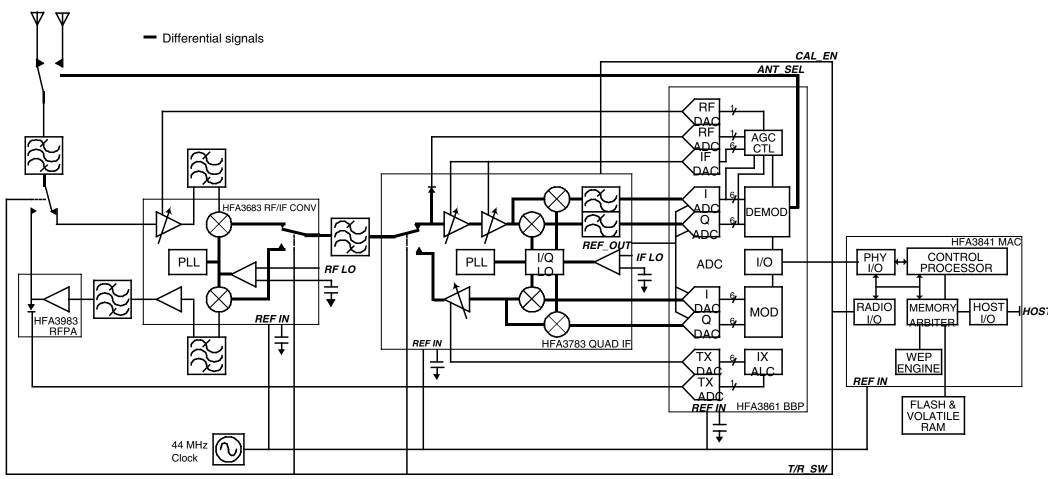


FIGURE 6.24 Intersil PRISM II® radio block diagram.

implementation uses five chips and has a total component count of approximately 200. The major chips include an RF converter that operates using an RF frequency of 2.400–2.484 GHz and an IF frequency of 374 MHz. A synthesizer is included in the IC with ceramic filters and a VCO provided at the IC inputs. A PA boosts the signal to approximately +20 dBm prior to final T/R switching, filtering, and any diversity switching. An IF IC converts the signal to baseband or modulates the transmit signal following filtering with a single SAW filter at 374 MHz. The MODEM IC implements the IEEE 802.11 CCK modulation with special circuits for multipath corrections. Finally a digital IC implements the media access controller (MAC) function for 11 Mbit data and interfaces with the computer/controller.

A brief description of the signal paths begins on the upper left that the dual antennas which may be selected for diversity with a command from the BaseBand Processor (BBP). The desired antenna is routed to a ceramic "roofing" filter to attenuate out-of–band signals and attenuate the 1.6 GHz image frequencies. Front-end losses with Diversity, T/R switches, and filters are typically 4 dB. The 3 dB NF LNA has a selectable high gain of +15 dB or low gain of –15dB controlled by the BBP depending upon signal level. The range of signals from –90 to –30 dBm uses the LNA high gain mode and –29 to –4 dBm (IEEE 802.1 maximum) uses low gain on the LNA. The first active Gilbert Cell mixer has a gain of +8 dB and NF of 9 dB. A differential 374 MHz SAW filter with 8 dB loss follows the RF converter. An IF demodulator then processes the –80 to –20 dBm signal with a linearized AGC stage prior to the final baseband mixer conversion. The signal is low pass filtered and passed to the Base Band Processor (BBP). Digital BBP processing includes I/Q A/D, interpolating buffers with digital NCO phase locked carrier recovery and Rake receiver/Equalizer processing. Finally the received data is processed with the MAC CPU in accordance with the IEEE 802.11 protocol.

The transmit function begins with data to the MAC being processed for BBP modulation. The data modulator includes a self-synchronizing scrambler function that removes the periodicity of the short 11 chip Barker sequence. A BBP transmit Data Formatter and Spreading Table provides the PN spreading function. A digital filter then reduces the DQPSK –13 dB Side Lobe Level (SLL) to approximately –45 dBc. This digital baseband waveform is then sent to the BBP I and Q D/A converters. The IF chip's modulator upconverts these signals to the 374 MHz IF using an active Gilbert Cell mixer and combiner. This modulated signal is then Automatic Level Controlled (ALC'd) using a feedback algorithm with the HFA3983 PA's detector to maintain the IEEE 802.11 30 dBc SLL. This IF signal is filtered by the same SAW as used in the receive path. A final RF upconversion from 374 MHz to 2.4 GHz is performed by the RF converter using the integrated on-chip synthesizer. The output level of approximately –5 dBm is filtered for image and harmonics and applied to the PA and finally the T/R and Diversity Switches. Schematics, application notes, and additional details for this example design may be obtained at www.Intersil.com.

7

IEEE 802.11g Higher Data Rates in the 2.4 GHz Band

Jim Paviol
RF Consultant

Al Petrick
Jones-Petrick and Associates

Bob O'Hara
Cisco Systems

7.1 Introduction to IEEE 802.11g

Increasing data rate demand, following the 802.11b amendments for 11 Mb/s, in Wi-Fi networking for enterprise and home networking was triggered by applications such as streaming video and hot-spot deployments. The success of the 2.4 GHz band Wi-Fi market and the relative higher 5 GHz band difficulties of 802.11a with its high 54 Mb/s OFDM data rate led to the consideration of applying this modulation technology to the lower frequency 2.4 GHz band.

In 2001, IEEE 802.11 working group examined applying OFDM modulation, similar to 802.11a, in the 2.4 GHz band. FCC approval was a major concern with the smaller spectrum size and competing services. The IEEE working group decided to proceed in 2001 when the FCC indicated that their concerns were addressed. Up until May 2001 only 802.11b Wi-Fi devices using Barker word spreading, CCK, or packet binary convolutional code (PBCC) modulation were allowed in this band. In June 2003, 802.11g was ratified as the new physical layer extension that enabled 1–54 Mb/s data rates in the 2.4 GHz band using the existing MAC layer and compatible 802.11b 11 Mb/s CCK (complementary code keying) modulation.

7.2 Network Deployment and Users

The 802.11g WLAN network is designed to support three types of scenarios. These are (a) Legacy Wi-Fi 802.11b—CCK, (b) Mixed-Mode 802.11b—CCK/802.11g—OFDM, and (c) New Green Field—802.11g OFDM installations (see Figure 7.1). Legacy Wi-Fi devices receive and detect CCK signals only without

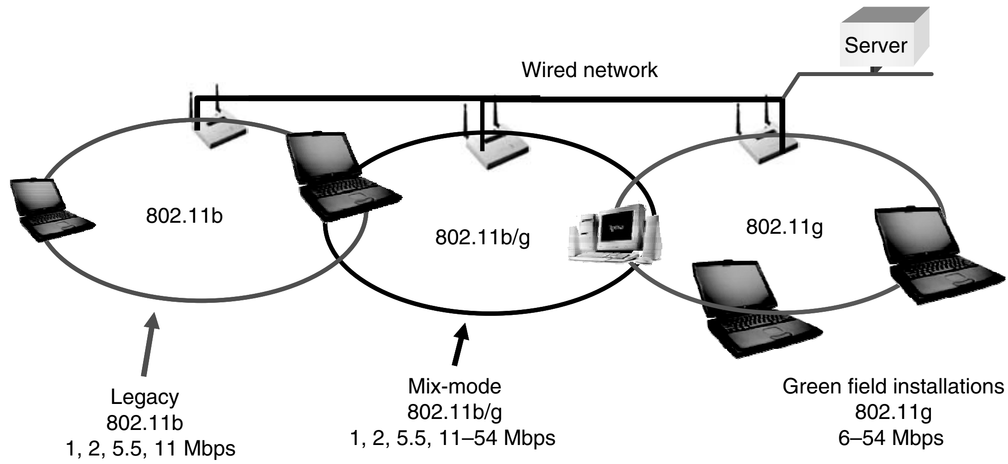


FIGURE 7.1 802.11g system deployment.

any OFDM signals. Initial deployments of 802.11g systems operate as mixed-mode networks with legacy 802.11b CCK—Wi-Fi Stations and 802.11g (CCK and OFDM) coexisting and interoperating with 802.11g access points. The mixed-mode systems fall back to CCK at <11 Mb/s data rates. New "green" field installations, where all Stations are 802.11g capable of CCK or OFDM operation, obtain up to the 54 Mb/s with through puts similar to the 802.11a 5 GHz systems but with improved 2 GHz frequency band propagation.

7.3 Mandatory and Optional Modes of Operation

802.11b CCK and 802.11a OFDM signaling are mandatory for the 802.11g physical layer (see Table 7.1). Each Station shall have three preamble and PHY layer convergence procedure (PLCP) header PHY protocal data unit (PPDU) formats. The mandatory data rates for transmitting and receiving data payloads are 1, 2 Mb/s for DSSS; 5.5 and 11 Mb/s for CCK; and 6, 12, and 24 Mb/s for OFDM. For DSSS and CCK it is mandatory for all Stations and access points to lock the Transmit center frequency and Symbol clock frequency to the same reference oscillator. The optional data rates supported for OFDM are 9, 18, 36, 48, and 54 Mb/s. Although 54 Mb/s is an optional data rate for compliance to 802.11g, it is mandatory for 802.11g devices obtaining Wi-Fi certification. For most applications such as the transmission of MPEG-2 video using wireless Standard Definition TV, a data rate of 24 Mb/s is sufficient. Other mandatory features for CCK include (a) support of Short Preamble, (b) maximum RF input signal limited to -20 dBm, and (c) Transmit center frequency and Symbol clock frequency must be locked to the same reference oscillator. Finally, the CCA mechanism must detect all mandatory Sync symbols.

7.4 Optional Modes of Operation

802.11g specifies CCK-OFDM and PBCC as optional modes of operation. CCK-OFDM and PBCC are allowed by the standard but are not required to implement 802.11g Stations or Access Points. CCK-OFDM is referred to as a hybrid waveform that combines CCK (single carrier) and OFDM (multicarrier) modulations for transmission of 802.11g packets. The preamble and PLCP header are transmitted using CCK, and the PHY service data unit (PSDU) payload is transmitted using OFDM. The short and long preamble used in 802.11b are specified for CCK-OFDM transmission (see Figures 7.2 and 7.3). This allows legacy 802.11b Wi-Fi devices to receive and detect the preamble of the CCK portion of the CCK-OFDM

TABLE 7.1 Mandatory Modes of Operation

Function	Specification	Notes
Carrier frequency	2.400–2.4835 GHz	Mandatory—worldwide
Channel mask	CCK , OFDM	Same as 802.11a and 802.11b
Channel spacing	25 MHz channel spacing	Same as 802.11b
Modulation type	DSSS, CCK, and OFDM	Mandatory (CCK Short and Long Preamble)
Data rates	1, 2, 5.5, 6, 11, 12, and 24 Mb/s	Mandatory (1, 2 Mb/s for DSSS, 11 Mb/s for CCK, 24 Mb/s for OFDM)
Data rates	9, 18, 36, 48, and 54 Mb/s	Optional
CCA detection	CCK and OFDM Preamble Sync Symbol	Mandatory
Transmitter clock lock	DSSS-CCK	Mandatory (Center frequency and Symbol clock)

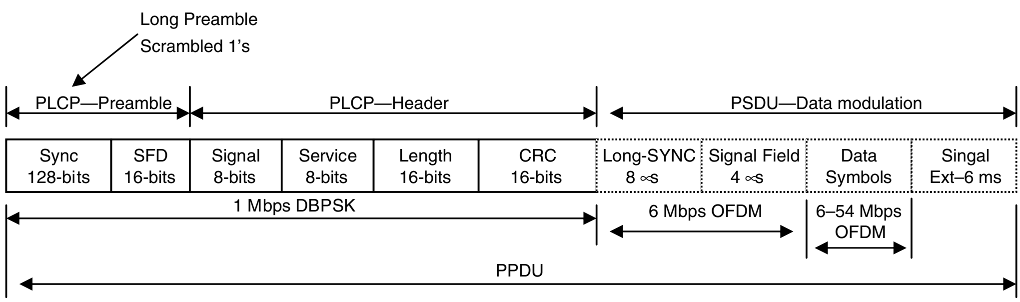


FIGURE 7.2 CCK–OFDM Hybrid—Long preamble.

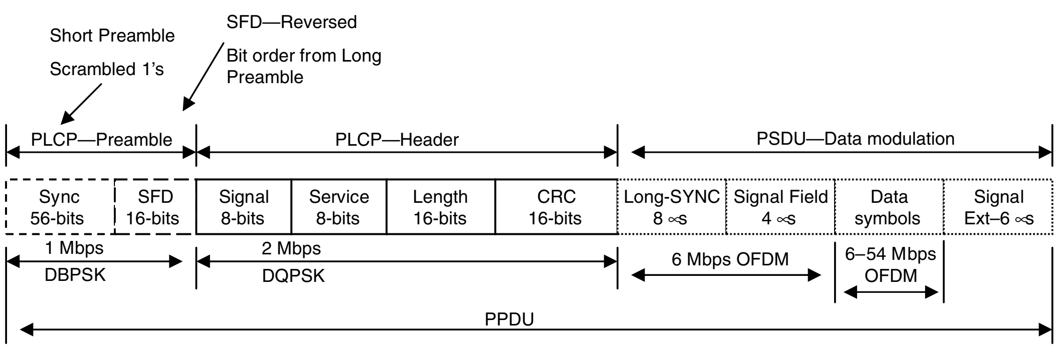


FIGURE 7.3 CCK–OFDM Hybrid—Short preamble.

transmission and defer the medium to prevent the possibility of collisions. Therefore, coexistence between CCK-OFDM and legacy 802.11b Wi-Fi devices on the channel is not a problem.

During transmission of a long preamble, the PLCP is transmitted at 1 Mb/s using DBPSK modulation for the PPDU and the PSDU payload transmits the long OFDM symbols and Signal Field at 6 Mb/s and the data symbols at the payload rate specified in the Signal Field (6–54 Mb/s). During transmission of the Short Preamble in the PPDU, 56-bit Sync field and Start of Frame Delimiter (SFD) field are transmitted at 1 Mb/s DBPSK modulated and the PSDU is transmitted at the OFDM rates (6–54 Mb/s).

PBCC is the second optional mode of operation. PBCC is also an option in the legacy 802.11b physical layer specification. Like CCK-OFDM, PBCC is considered a hybrid waveform. PBCC employs a complex signal constellation 8-PSK structured using a convolution coding technique that builds on the transmission of CCK preamble. PBCC transmits the preamble and PLCP header using CCK and payload is transmitted using PBCC. PBCC supports the basic rates: 1, 2, 5.5, and 11 Mb/s and employs a code structure that extends the data rates to 22 and 33 Mb/s.

TABLE 7.2 Service Field Revised for 802.11g

b0	b1	b2	b3	b4	b5	b6	b7
Reserved	Reserved	Clock lock bit 1 = locked	Modulation 0 = CCK 1 = CCK–PBCC	Reserved	Length extension PBCC 22-33	Length extension PBCC 22-33	Length extension

7.5 PPDU Formats

All 802.11g Stations must support three types of preamble and PLCP headers. They are as follows:

1. Long Preamble and PLCP header (see Table 7.2 for revised Service Field)
2. Short Preamble and PLCP header
3. OFDM Preamble and PLCP header

Service Field: The Service Field bit positions are shown in Table 7.2. Data bits (b0), (b1), and (b4) are reserved bits and always set to logic 0. The Service field uses 3-bits of the reserved 8-bits for 802.11b. Data bit (b2) determines whether the Transmit frequency and Symbol frequency clocks use the same local oscillator. Setting this bit to logic 1 is mandatory for all 802.11g Stations and Access Points. Data bit (b3) indicates whether CCK or PPBC is used. Data bit (b7) is a bit extension used in conjunction with the Length field to calculate the duration of the PSDU, in microseconds for CCK if data bits (b3), (b5), and (b6) are set to logic 0. If data bit (b3) is set to logic 1 then data bits (b5), (b6), and (b7) are used to resolve data field length resolution in microseconds for PBCC 22 Mb/s and PBCC 33 Mb/s optional modes. To select PBCC-22 the Service Field is set to DCh, for PBCC −33 Mb/s set to 21 h, and for all data rates DSSS-OFDM set to 1Eh. The data bits in the Service field are transmitted LSB first.

7.6 Operating Channels

The spectral mask for 802.11g in the 2.4 GHz frequency utilizes the masks defined for 802.11b—CCK and 802.11a—OFDM. The transmit masks are designed to allow three noninterfering channels spaced 25 MHz apart over the 2.4 GHz frequency band. This allows CCK and OFDM Stations and Access Points to coexist spectrally with legacy 802.11b Wi-Fi networks and provides a seamless upgrade path to employ 802.11g. Figure 7.4 illustrates the channel mask and spacing using channels 1, 6, and 11 for an 802.11g system using OFDM. The 3 dB channel bandwidth of the OFDM channel is 20 MHz. Therefore using the legacy channnelization plan, with 25 MHz channel spacing adjacent channel interference is not an issue. However when operating at the lower and upper band edges of the 2.4 GHz frequency band, additional filtering or back off in the RF Transmit power may be necessary to conform with the local regulatory requirements. Figure 7.5 shows channel 6 operating as an 802.11b–CCK network. In this example, channels 1 and 11 are operating as OFDM networks illustrating the capability of mixed-mode operation.

7.7 Operation of 802.11g—CSMA/CA and CCA

Just as with the existing 802.11a and 802.11b ratified amendments 802.11g uses the listen-before-talk mechanism built into the MAC protocol to transmit packets over the medium. This mechanism is well known as CSMA/CA where by which a Station must listen and sniff the medium for a clear channel before transmitting packets. This is performed by using the clear channel assessment (CCA) mechanism. For 802.11g, all Stations must support the three preamble and header mandatory formats as previously described. This ensures that all implementations are capable of receiving and detecting DSSS, CCK, and

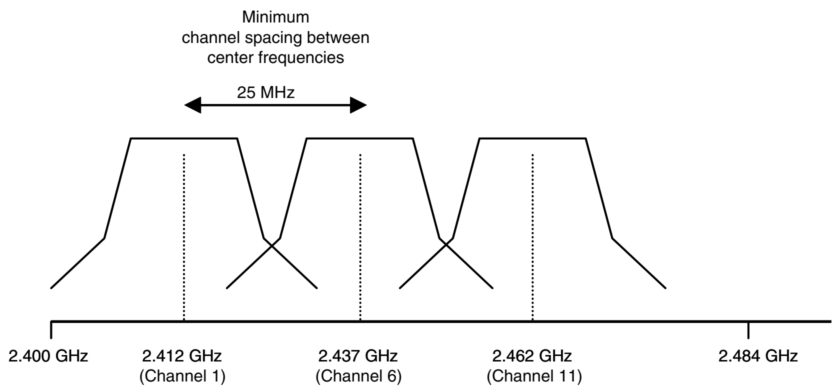


FIGURE 7.4 Channel mask spacing for OFDM operation.

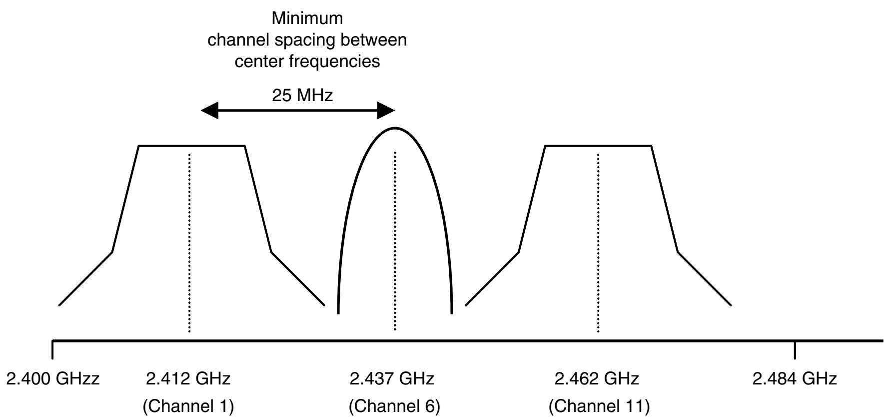


FIGURE 7.5 Channel mask spacing for simultaneous operation of OFDM and CCK on adjacent channels.

OFDM signals, and the CCA mechanism must be capable of detecting a medium busy for each PLCP and PPDU type.

CCA detection is based on using the combination of energy detect (ED) and carrier sense (CS), and CS in the receiver must be capable of detecting all Barker and OFDM synchronization symbols in the preamble. Without this capability, both packet collisions and lack of coexistence between CCK and OFDM Stations would occur, and potentially severally impair the medium. In a given BSS, which consists of an access point and all associated Stations must listen for transmissions to determine if the channel is idle. If the channel is idle where no minimal energy or carrier is detected then an internal back off timer begins to count down. Once the timer expires the Station begins transmission. To ensure that the possibility of collisions from other Stations in the BSS is minimized, other mechanisms are employed in the back off timer to randomize the duration of the timer. However on the other hand if the receiver's CCA mechanism is triggered by ED, which detects the minimum signal power threshold equal to or greater than −76 dBm and if the minimum ED is met and the receiver demodulates a DSSS or OFDM preamble, CCA is reported as medium busy and transmission of the Station requesting to transmit information waits for a clear medium using the internal back off timer.

Under normal operation, all Stations sharing the medium can hear and detect each other. However in the case of the hidden node problem, CCK devices cannot hear OFDM transmissions. To overcome this problem the RTS/CTS mechanism built in the standard is used for mixed-mode deployments where legacy CCK and OFDM exists in a BSS. When the CTS/RTS mechanism is employed, most Stations in a BSS will hear and detect RTS and all Stations will hear the CTS from the access point. During a RTS/CTS exchange each Station receives information on how long the CCK or OFDM packet including the ACK

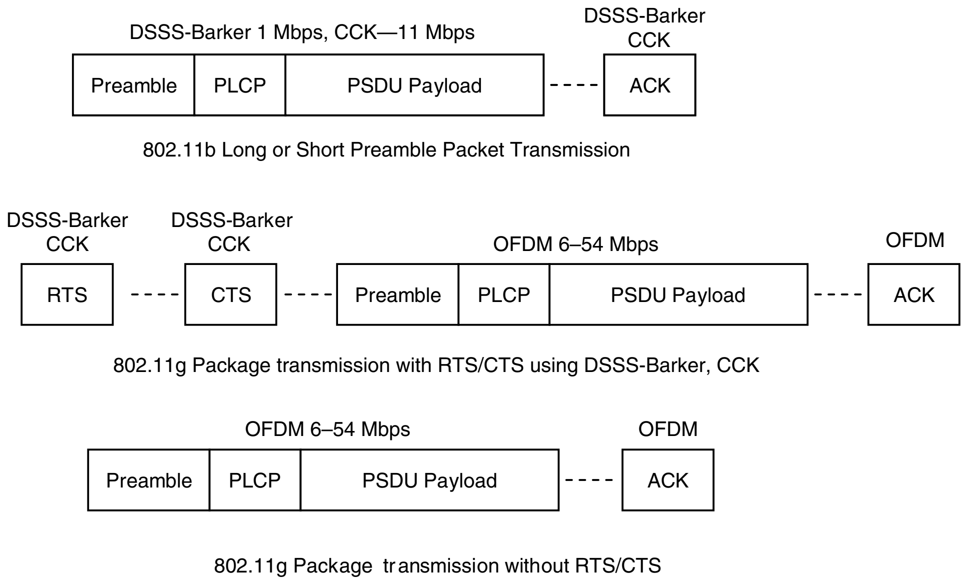


FIGURE 7.6 802.11g CCK–OFDM packet transmission.

TABLE 7.3 Key System Parameters

Function	CCK	OFDM	Notes
Slot time	20 μsec	9 μsec	
CCA	15 μsec	4 μsec	Must detect all OFMD Sync symbols
Probability of detection—CCK	>99%	>90%	
Symbol frequency clock	±25 ppm	±25 ppm	Transmit and carrier frequency are derived from the same clock for DSSS and CCK (Locked oscillators)
Transmit center frequency accuracy	±25 ppm	±25 ppm	Transmit and carrier frequency are derived from the same clock for DSSS and CCK (Locked oscillators)
Packet error rate	10%	10%	Measured with PSDU packet length of 1000 bytes
SIFS	10 μsec	10 μsec	
Receiver input level	−20 dBm	−20 dBm	Maximum

transmission will be. The NAV, which is an internal timer, in each Station is set to have the same duration as the OFDM packet exchange. The NAV timer and the internal back off timer operate together as part of the CCA virtual CS mechanism. If the NAV and back off timer expires and no energy or CS is detected, the channel is claimed to be idle and the Station begins to contend for the medium access. It is this operation that allows the coexistence between CCK and OFDM radios in a BSS (Figure 7.6). This figure shows the packet exchange for CTS/RTS with CCK and OFDM transmission with ACK exchanges. It is also noted that all Stations and the Access Point in a BSS 802.11g network must be capable of gear shifting back in data rate and operate as a legacy 802.11b device.

7.8 Key System Specifications

Table 7.3 lists a number of key specifications required for CCK and OFDM operation for 802.11g. The specifications were derived from the 802.11b—CCK and 802.11a—OFDM physical layer and MAC layer protocol specifications.

8

Wireless Personal Area Network Communications, an Application Overview

Ian C. Gifford
Freescale Semiconductor, Inc.

8.1 Introduction to WPAN Communications

The past five years, since the last printing of the handbook, has seen an explosion of applications in the wireless personal area network (WPAN) space. In the first edition of this chapter on WPANs we talked about the then emerging Bluetooth™ wireless technology; today Bluetooth is an everyday word, specifically "Bluetooth technology is shipping today at a rate of over 13 million devices per week and has an installed base of 1 billion devices. Improved pairing and lower power consumption will enhance the Bluetooth experience for all users and speed the technology's growth to the next billion Bluetooth devices."* This rapid growth is being driven not only by increasing attach rates in the important mobile phone segment but also newly emerging segments; for example, consumers are having to choose if they want their Bluetooth-enabled automobile with a hands-free car kit and a 2 year cellular service contract! WPAN communication is also emerging in consumer electronics, that is, as fashion enabled accessories, toys, medical enabled devices as well as sensors. To be sure there are other WPAN wireless technologies but far and away the Bluetooth wireless technology is leading the WPAN network space. Bluetooth is the über WPAN connectivity solution.

8.2 WPAN Market Drivers

The killer application for the Bluetooth is clearly the ubiquitous headsets we see in our everyday life, for example, a small "ear bud" or headset that may have a blue light glowing and the user talking into the ether. At first the person appears to be talking to themselves but then you notice the headset and you say to yourself "I want one of those!"

Also this year has seen a conscious effort to promote hands-free system legislation to increase and ensure safe phone handling in the car. This is a significant driver and will continue to drive Bluetooth adoption during the later half of this decade.

As expected the highly vertical slow adopter markets, that is, automotive, industrial, medical, and so on are looking at open WPAN specifications and specifically Bluetooth wireless technology as a viable solution versus proprietary wireless technologies. Medical applications such as wireless medical sensors, remote out patient monitoring, fitness and workout tracking, and so on.

Indirectly, the WPAN space is part of a larger democratization of the person, that is, the explosive growth of the Internet and social networks,—that is, individual user-generated content, such as found on MySpace,† YouTube,‡ and so on—is driving astronomical user-generated forecast for downloads and uploads during the later half of this decade. The WPAN is on the roadmap for the evolving devices that will facilitate the personal content generated to populate these web sites. Examples are simple headsets, via Bluetooth, for stereo audio to more complex WPAN technologies, for example, streaming video applications, online gaming, and so forth.

The outlook for WPANs is very strong and does not appear to be abating. In Figure 8.1, the WPAN as the source of user-generated internet content will be a significant impact to the Internet, Internet-based markets, alternate content sourcing and distribution models, and so on.

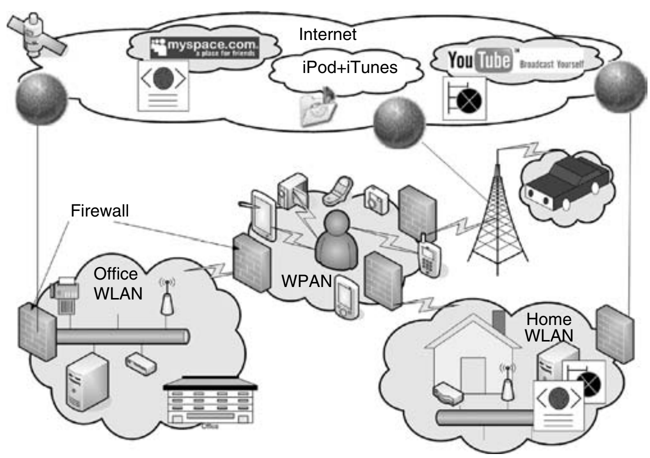


FIGURE 8.1 WPAN as the source of user-generated internet content.

* http://www.bluetooth.com/Bluetooth/Press/SIG/BLUETOOTH_SIG_IMPROVES_USER_EXPERIENCE.htm
† http://www.myspace.com/
‡ http://www.youtube.com/

The emergence of the WPAN as the center of this content multiplier will drive demand even higher for WPAN-based wireless technologies. The installed base of Bluetooth devices reached 1 billion in early November of 2006. Weekly shipments of the wireless devices continue at a pace of 13 million per week. The expectation in 2007 is shipping 2 billion devices in one year in 2010.

In addition to the Bluetooth, market drivers are the WPAN-based wireless sensor markets found in the ZigBee Alliance and IEEE Std 802.15.4™-2006 as well as the newly emerging Wibree radio technology too. These sensor networks will be significant drivers to the WPAN market as sensors are vertically adopted by less buyers versus horizontally by more buyers.

However, the primary market driver in the WPAN market will be the Bluetooth wireless technologies and their umbrella technologies too.

8.3 Introduction to Bluetooth

Bluetooth is the name for the specification that has been described by the IEEE standard 802.15.1™-2005 for short-range wireless connections between desktop and notebook computers, handhelds, personal digital assistants, mobile phones, camera phones, printers, digital cameras, headsets, keyboards, mouse, and other devices. Bluetooth uses a globally available frequency band (2.4 GHz) for worldwide compatibility. Considered an alternative to the IrDA (Infrared Data Association) standard, it is commonly used to connect devices within a 10 m range at transfer rates less than 3 Mbps. Its features include the following:*

- Bluetooth wireless technology is geared toward voice and data applications.
- Bluetooth wireless technology operates in the unlicensed 2.4 GHz spectrum.
- Bluetooth wireless technology can operate over a distance of 10 m or 100 m depending on the Bluetooth device class. The peak data rate with enhanced data rate (EDR) is 3 Mbps.
- Bluetooth wireless technology is able to penetrate solid objects.
- Bluetooth technology is omnidirectional and does not require line-of-sight positioning of connected devices.
- Security has always been and continues to be a priority in the development of the Bluetooth specification. The Bluetooth specification allows for three modes of security.
- The cost of Bluetooth chips is under $3.

8.4 Current Status of Bluetooth

The current status, as of this writing, of the Bluetooth core specifications is that the Bluetooth v2.1 + EDR released on March 27, 2007 is the approved specification. On the basis of the groups' publication cycle and their explosive base it follows that an update should be expected, that is, v3.0 in 2008 or later. Minor prototyping specifications may be released as interim updates, that is, errata, new or revised profiles, and so on.

The year 2005 was a banner year for the Bluetooth SIG, which saw the emergence of the automotive and stereo audio industries to help drive the current specification adoption rate. In 2007, with the release of Bluetooth v2.1 + EDR the core specification was updated with Improved Pairing, Lower Power Consumption, and improved security too.

The Bluetooth specifications can be freely downloaded† from the bluetooth.com web site. There are various profile specification that are vertical slices or system end-to-end solutions and are required reading. However, the Bluetooth v2.0 + EDR core specification has the most deployment and is a very dense technical specification containing 1230 pages or 13 MB PDF file. There are four volumes:

- Volume 0 Master Table of Contents & Compliance Requirements
- Volume 1 Architecture & Terminology Overview

* http://www.bluetooth.com/Bluetooth/Learn/Technology/Compare/
† http://www.bluetooth.com/Bluetooth/Learn/Technology/Specifications/

- Volume 2 Core System Package [Controller volume]
- Volume 3 Core System Package [Host volume]

Each of the Bluetooth volumes is a self-contained "book," which has its own table of contents. Each volume covers one or more parts (A, B, etc.), each part can be viewed independently and has its own table of contents. Volume 0 contains a high level table of contents for the remaining volumes and a Compliance Requirements document. Volume 1 is very helpful in describing the high level architecture of the Bluetooth wireless technology. Volume 2 has the key "controller" parts, that is, where the radio, baseband, link manager, host controller interface (HCI), and so on are described. Volume 3 has the key host parts, that is, where the higher layer "host" parts, that is, L2CAP, service discovery protocol (SDP), gernic access profile (GAP), and so on are described.

In Figure 8.2 the core system architecture is depicted. The core system provides basic services:

- Device control services that modify the behavior and modes of a device.
- Transport control services that create, modify, and release traffic bearers.
- Data services that are used to submit data for transmission over traffic bearers.

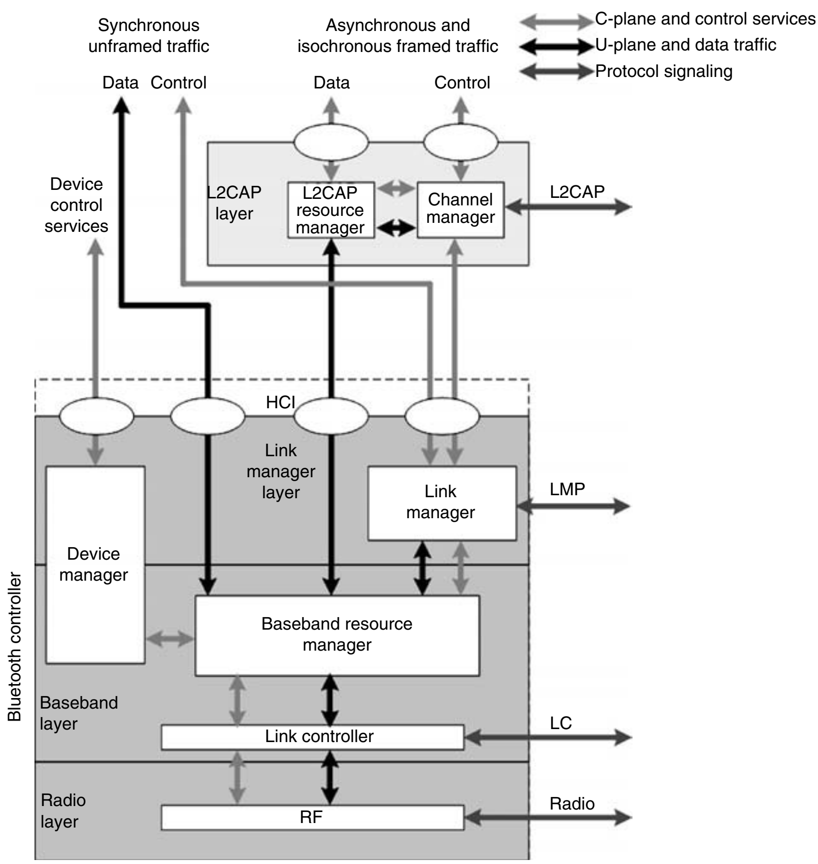


FIGURE 8.2 Bluetooth core system architecture. (From BSIG and IEEE.)

In summary, the Bluetooth core system consists of an RF transceiver, baseband, and protocol stack. The HCI provides a command interface to the baseband controller and link manager, and access to configuration parameters. This interface provides a uniform method of accessing the Bluetooth baseband capabilities. The system offers services that enable the connection of devices and the exchange of a variety of classes of data between these devices.

8.5 Future Status of Bluetooth

The Bluetooth wireless technology adoption is exciting and with 1 billion devices shipped in November of 2006 makes this technology the leading wireless technology in the WPAN space.

The new Bluetooth SIG provides the following vision for the group "The potential for and momentum behind Bluetooth wireless technology is increasing steadily as consumer awareness around the world continues to rise and new industries like the health and fitness market embrace the advantages of Bluetooth technology in exciting new applications. My goals include the adoption of an alternative MAC/PHY to address high speed Bluetooth applications, increasing the level of interoperability and ease of use of Bluetooth enabled products, and expanding the adoption of Bluetooth wireless technology."*

The "adoption of an alternative MAC/PHY to address high speed Bluetooth applications" refers to the recent announcement of the Bluetooth SIG and the WiMedia Alliance "WiMedia Alliance multiband orthogonal frequency division multiplexing (MB-OFDM) version of ultra wideband (UWB) for integration with current Bluetooth wireless technology."† This complementary wireless technology is the first of a series of alternate wireless technologies that have been identified; the Bluetooth SIG is planning to work with the following wireless technologies:‡

- UWB is a wireless technology with data speeds of 100 Mbps to over 2 Gbps that transmits across a wide frequency range at once. Potential scenarios for collaboration between UWB and Bluetooth technology would include short-range file transfer, streaming high-quality video between portable devices, and so on. Bluetooth specifically will focus on UWB radios that operate above 6 GHz (mobile devices go everywhere), aggressive power requirements, and roadmap to single chip solutions.
- IEEE 802.11™ also known as Wi-Fi® described by the IEEE standard 802.11™ is a wireless technology developed for WLAN applications connecting computers, phones, and consumer electronics to the Internet at speeds exceeding 50 Mbps when in range of an access point; draft P802.11a/n targets 100 Mbps at the MAC SAP. Potential scenarios for collaboration between 802.11™ and Bluetooth technology would include home gateways and media centers that would include both technologies. In addition, with both 802 and Bluetooth technologies going into mobile devices, it is important to work together to identify ways the technologies can optimize space and power consumption.
- NFC (near field communication) is a wireless technology that covers a short distance, a few centimeters or less. NFC is optimized for secure communication between devices with minimal user configuration. A possible collaboration between NFC and Bluetooth technology could occur in device pairing—the identification process between two Bluetooth-enabled products first connecting.
- Wibree radio technology is a wireless technology that Nokia has developed with a few partners. A possible collaboration between Wibree and Bluetooth technology could occur in mobile phones, multimedia computers, and PCs as a dual-mode radio that provides an add-on functionality to the Bluetooth wireless technology where it networks and/or controls "sensor radios," and so on.

The following is an example of two of these "umbrella technologies" for the future Bluetooth wireless technologies.

* http://www.bluetooth.com/Bluetooth/Press/SIG/BLUETOOTH_SIG_NAMES_DR_JOHN_R_BARR_MOTOROLA_CHAIRMAN_OF_THE_BOARD.htm

† http://www.bluetooth.com/Bluetooth/Press/SIG/BLUETOOTH_SIG_SELECTS_WIMEDIA_ALLIANCE_ULTRAWIDEBAND_TECHNOLOGY_FOR_HIGH_SPEED_BLUETOOTH_APPLICATION.htm

‡ http://www.bluetooth.com/Bluetooth/Press/SIG/Bluetooth_SIG_Unites_Wireless_Under_One_Umbrella.htm

8.6 Bluetooth + UWB AMP

With the addition of UWB, the Bluetooth + UWB enabled device will be able to provide high-speed applications realizing >3 Mbps, for example, 24, 50, 100 Mbps and beyond. The addressable applications for the Bluetooth wireless technologies increase significantly with the alternate high speed channel.

As an example during last years' 2006 Bluetooth SIG Annual Meeting there was a demonstration* of a stream high definition (HD) video streaming using the Bluetooth video distribution profile (VDP) over UWB radios while full Bluetooth wireless technology interoperability is maintained, while video is being streamed and files are being transferred, a picture is pushed from a mobile phone using the Bluetooth Object Push Profile.

Additional use cases will be developed to take advantage of this higher speed channel in the Bluetooth WPAN.

8.7 Bluetooth and Wibree

The WPAN Bluetooth revolution continues just outside of the special interest group too, for example, on October 3, 2006 Nokia introduced Wibree technology as an open industry initiative. The focus of Wibree is on ultra low-power devices such as watches and sensors and is a complement to Bluetooth and not necessarily a substitute. "Wibree is implemented either as stand-alone chip or as Bluetooth-Wibree dual-mode chip."† With the introduction of the proprietary Wibree technology application providers can choose to implement the dual-mode alternative and take advantage of existing Bluetooth wireless technology on their platform, for example, dual mode for mobile devices, mobile phones, and so on, or stand-alone application providers can target watches, sensors, and so on.

Specifically, the "Wibree radio specification enables dual-mode implementations to reuse Bluetooth RF part but also to guarantee ultra low power consumption for devices with embedded stand-alone implementation of the Wibree specification. Wibree operates in 2.4 GHz ISM band with physical layer bit rate of 1 Mbps and provides link distance of 5–10 meters."‡ The Wibree transmission will operate, discover, and so forth in between Bluetooth wireless technology transmissions. It follows that Wibree profiles will be developed to address the dual-mode and stand-alone implementations.

The Wibree radio technology appears to be addressing the same area that IEEE Std 802.15.4 and ZigBee Alliance are addressing.

8.8 IEEE WPAN Activities

The IEEE 802.15 Working Group (WG) for WPANs§ continues to develop consensus standards, recommended practices, and guides for the WPAN area. Here is a current listing of their free** published documents:

- IEEE 802.15.1™—2005—Standard for Bluetooth v1.2
- IEEE 802.15.2™—2003—Recommended Practice Coexistence of Wireless Personal Area Networks with Other Wireless Devices Operating in Unlicensed Frequency Bands
- IEEE 802.15.3™—2003—Standard MAC and PHY Specifications for High Rate WPANs
- IEEE 802.15.3B™—2005—Standard MAC and PHY Specifications for High Rate WPANs Amendment 1: MAC Sublayer

* http://bluetooth.com/Bluetooth/Press/News/Bluetooth_Wireless_Technology_over_UWB_Demonstrated_at_Bluetooth_SIG_All_Hands_Meeting.htm

† http://www.wibree.com/press/

‡ http://www.wibree.com/technology/

§ http://ieee802.org/15

** http://standards.ieee.org/getieee802/

- IEEE 802.15.4™—2006—Standard MAC and PHY Specifications for Low Rate WPANs (includes 15.4B work)

These WPAN standards are part of the larger IEEE 802® standards development organization. Figure 8.3 describes the family of standards for local and metropolitan area networks. The relationship between the

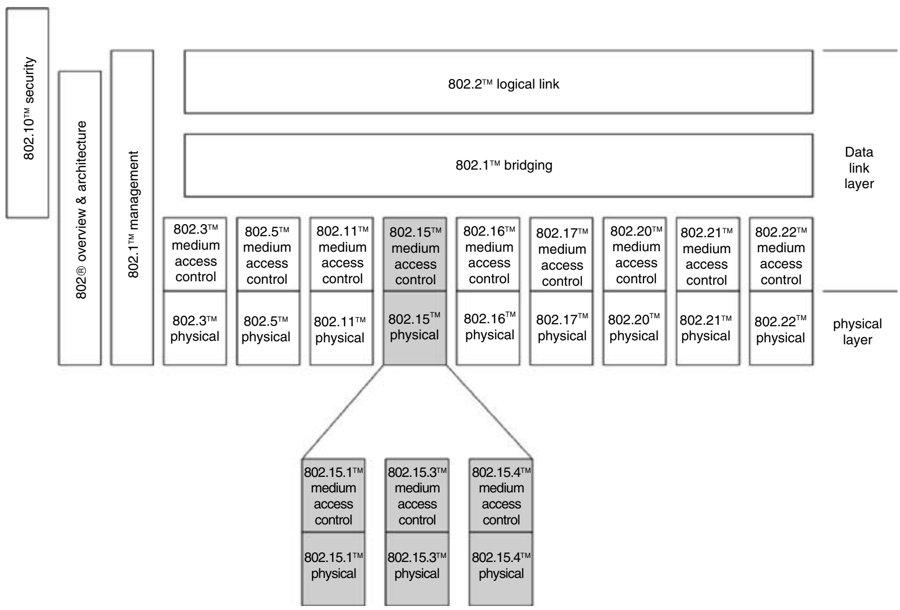


FIGURE 8.3 IEEE 802 family of standards.

WPAN standards and other members of the family is shown below. The numbers in the figure refer to IEEE standards numbers. The WPAN standards focus only on the medium access control (MAC) and physical layer (PHY) of the ISO OSI seven layer model. In the development of the IEEE WPAN standards there has been unique MAC and PHY specifications created to address the unique applications and use cases that are being addressed by this consensus standards development organization.

Specifically, there are numerous IEEE-based WPAN MACs and PHYs that have been created and more are being developed to continuously provide the optimum common air interface both from a geographic and/or use case based approach. The IEEE Std 802.15.1™-2005 has a single MAC and PHY combination based on the Bluetooth v1.2 core specification; the operating frequency is 2400–2483.5 MHz. See the Bluetooth SIG web site* for the latest specification.

The IEEE 802.15.3™-2003 provides a single MAC and PHY where the PHY provides a single carrier system that supports up to five modulation formats with coding at 11 Mbaud to achieve scalable data rates. Table 8.1 provides the formats, coding and data rates for the 802.15.3 PHY.

TABLE 8.1 IEEE Std 802.15.3-2003 PHY Rates

Modulation Type	Coding	Data Rate Mb/s
QPSK	8-State TCM	11
DQPSK	None	22
16-QAM	8-State TCM	33
32-QAM	8-State TCM	44
64-QAM	8-State TCM	55

* http://www.bluetooth.com/Bluetooth/Learn/Technology/Specifications/

The popular (some would say infamous) Task Group 3a focused on developing a higher speed alternate PHY and was very active from 2001 to 2005; however, the group polarized into two technology camps, that is, MB-OFDM and DS-UWB and the groups decision making stalled. The goal of this project was to provide an alternate PHY with a "data rate…high enough, 110 Mb/s or more, to satisfy an evolutionary set of consumer multimedia industry needs for WPAN communications."* There were a variety of technologies proposed but in the end the group focused on adopting a PHY based on UWB. In the fourth quarter of 2005 discussion took place and in January 2006 the Task Group 3a finally agreed on a motion to withdraw the alternate PHY project. Interestingly, the MB-OFDM PHY and MAC technologies were in a parallel standards track while it was an active IEEE project and the result was an approved (December 2005) European computer manufacturers association (ECMA) International set of standards, that is, standard ECMA-368† High Rate UWB PHY and MAC Standard and the companion standard ECMA-369‡ MAC-PHY Interface for ECMA-368. These ECMA standards are being used by industry groups, that is, WiMedia, USB-IF, and Bluetooth SIG to promulgate the MB-OFDM technology as a high-speed WPAN.

The IEEE 802.15.4-2003 standard was approved in May 2003 and was published in October of the same year. The IEEE 802.15.4-2006 standard was approved by the IEEE standards board in June of 2006. The 2006 revision is based on the Task Group 4b work which focused on refining the IEEE 802.15.4-2003 specification by clearing up ambiguities, resolving inconsistencies, and made specific extensions: a faster subGHz physical interface, extensions to support time synchronization, reduced unnecessary complexity, increased flexibility in security key usage, and considerations for newly available frequency allocations. Table 8.2 provides the available frequencies for IEEE 802.15.4.

TABLE 8.2 EEE Std 802.15.4 PHY Frequencies

PHY (MHz)	Frequency Band (MHz)	Spreading Parameters		Data Parameters		
		Chip Rate (kchip/s)	Modulation	Bit Rate (kb/s)	Symbol Rate (ksymbol/s)	Symbols
868/915	868–868.6	300	BPSK	20	20	Binary
	902–928	600	BPSK	40	40	Binary
2450	2400–2483.5	2000	O-QPSK	250	62.5	16-Ary Orthogonal

The IEEE 802.15 WG continues to meet six times per year to deliberate various work items. The following is short summary of the current work items as of this edition; it is recommended that for those that want more detail to go and attend the plenary and/or interim sessions.§

8.9 Task Group 3c—60 GHz WPANs

Task Group 3c is focused on developing a millimeter-wave based (57–64 GHz) alternate PHY that will provide very high data rates at a mandatory 2 Gbps over 4 channels in 7 GHz bandwidth; for emerging use cases such as high-speed internet access, streaming content download (video on demand, HDTV, home theater, etc.), real time streaming, and wireless data bus for cable replacement. An optional data rate in excess of 3 Gbps will be provided too.

* http://standards.ieee.org/board/nes/projects/802-15-3a.pdf
† http://www.ecma-international.org/publications/standards/Ecma-368.htm
‡ http://www.ecma-international.org/publications/standards/Ecma-369.htm
§ http://grouper.ieee.org/groups/802/15/pub/Meeting_Plan.html

WPANs based on a 60 GHz radio are unique in that they provide a *natural* WPAN due to the unique oxygen absorption properties* of the 60 GHz spectrum.

The TG3c recently approved five usage models also known as use cases that are briefly summarized below:

Usage models

1. Uncompressed Video Streaming: 3.56 Gbps
2. Multi Uncompressed Video Streaming: 1.75 Gbps
3. Office Desktop: 3.56 Gbps
4. Conference ad hoc: 1.75 Gbps
5. Kiosk file-downloading: 2.25 Gbps

Figure 8.4 is a Kiosk file-downloading use case depicted with the traditional actor and use case callouts.

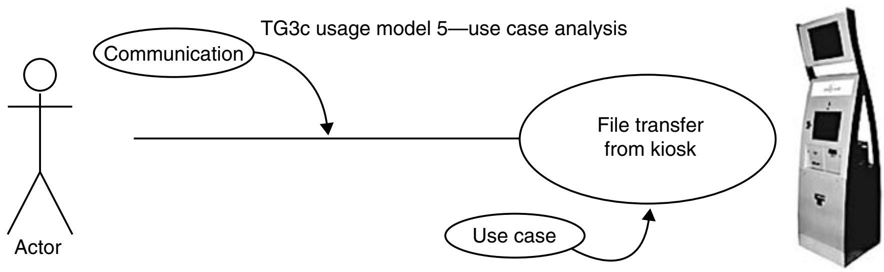


FIGURE 8.4 Project 802.15.3c use case—Kiosk file-downloading.

The person (actor) is in a store and wants to download a few movies into his/her mobile device. Ideally the person is 1–3 m from the Kiosk and has a line-of-sight, that is, unobstructed view of the Kiosk. There is a compatible radio transmitter/receiver in the Kiosk in a store as well as in the mobile device (cell phone, PDA, PMP, digital camera); ideally the device is placed next to the Kiosk, that is, the device is something the person is carrying. In addition, it is assumed that the device relative to the Kiosk is fixed during the task, that is, file transfer from Kiosk to the mobile device.

Figure 8.5 use case diagram depicts the Kiosk file-downloading use case with a scenario for the user. The background for this diagram is that the download speed is 2 Gbps and due to the extremely high speed the battery drain is diminished and the user experience is heightened.

This use case is just one example of this exciting new WPAN technology.

The Task Group 3c has received numerous contributions on the usage models and has developed summaries and posted them to their web site; please refer to IEEE 802.15† for further information on these models and application scenarios.

The Task Group 3c is actively working; they continue to update and refine their channel modeling work for the various use cases. Sixteen (16) proposals were accepted and presented at Montreal meeting in May 2007. Updated/merged proposals will be presented in the next meeting in July 2007. Their current estimate is to publish a standard in the second half of 2008.

---

* "The clear atmosphere (dry air mass) represents a unique filter over the 45 to 76 GHz range with frequency-dependent absorption properties caused by the microwave spectrum of oxygen not found at any lower frequency. The basic information necessary for determining molecular line, band, and continuum absorption is contained in a transfer function model. Close to sea level, an unstructured band exists, centered at 60 GHz with a half-width of 8 GHz and a pressure-dependent intensity that is corrected for pressure induced line overlap effects." Liebe, H. J.

† http://ieee802.org/15

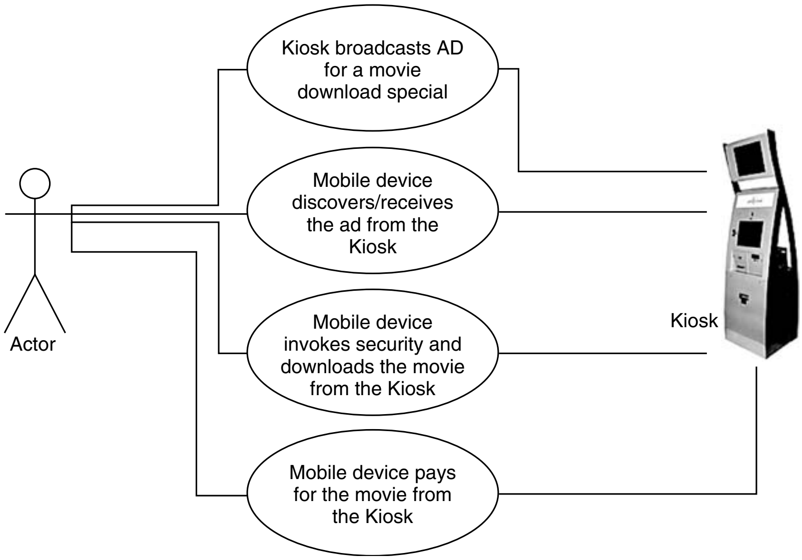


FIGURE 8.5 P802.15.3c use case diagram—Kiosk file-downloading.

8.10 Task Group 4a—Low-Rate WPANs

The Task Group 4a—low-rate WPAN is nearly approved and will be published shortly; their goal is to provide an international standard for an ultra low complexity, ultra low cost, ultra low power consumption alternate PHY for 802.15.4 (comparable to the goals for 802.15.4–2003). The draft standard is in IEEE-SA sponsor ballot and is expected to be published in the first half of 2007.

This amendment specifies two distinct PHYs: a UWB PHY at frequencies of 3–5, 6–10, and less than 1 GHz and a Chirp Spread Spectrum (CSS) PHY at 2450 MHz; CSS radios use a frequency-modulated pulse. The UWB PHY supports an over-the-air mandatory data rate of 842 kb/s, with optional data rates of 105 kb/s, 3.37 Mb/s, 13.48 Mb/s, and 26.95 Mb/s; the CSS PHY supports an over-the-air data rate of 1000 kb/s and optionally 250 kb/s. The PHY chosen data rate depends on local regulations, application, and user preference.

The UWB LR-WPAN specifications are designed to provide robust performance for LR-WPAN applications while leveraging the unique capability of UWB waveforms to support precision ranging between devices. The UWB PHY-layer design is intended to make use of the wide bands of spectrum being made available for UWB operation around the world. This spectrum, combined with advances in low-cost and low-power process technology, enables the implementation of LR-WPAN devices that can provide enhanced resistance to multipath fading for robust performance with very low transmit power.

The CSS 2450 MHz LR-WPAN specifications are designed to provide robust performance for LR-WPAN applications while leveraging the unique capability of CSS waveforms to support long-range links or to support links to mobile devices moving at high speeds. The CSS PHY-layer is intended to take advantage of the global availability of the 2450 MHz band due to favorable regulations based on preceding WPAN specifications, both indoors and out, while offering enhanced robustness, range and mobility. The properties of CSS enable LR-WPAN devices that can provide enhanced immunity to multipath fading and extended range for robust performance with very low transmit power. Two data rates are specified for CSS in order to offer the implementer the flexibility to select the rate and properties best suited for their application, as the following guidelines illustrate

- The lower/coded rate would be appropriate in quiet AWGN and high multipath environments.
- The higher rate would be appropriate for low-energy consumption and burst interference environments.

8.11 Task Group 5—WPAN Mesh

An 802.15 mesh network is a WPAN that employs one of two connection arrangements, full mesh topology or partial mesh topology. In the full mesh topology, each node is connected directly to each of the others. In the partial mesh topology, some nodes are connected to all the others, but some of the nodes are connected only to those other nodes with which they exchange the most data. Mesh networks have the capability to provide

- Extension of network coverage without increasing transmit power or receive sensitivity
- Enhanced reliability via route redundancy
- Easier network configuration
- Better device battery life due to fewer retransmissions

TG5 is still soliciting contributions to complement its baseline draft standard. In addition, they are actively meeting with other IEEE Mesh Groups, that is, 802.11s and 802.16j to share development ideas and contributions.

8.12 Study Group 4c—Low-Rate WPANs

The Study Group 4c is to looking into how to incorporate CWPAN (Chinese WPAN) standard into IEEE802.15.4; to accommodate the local regulations.

8.13 Task Group 4d—Low-Rate WPANs

The Task Group 4d is looking at amending IEEE Std 802.15.4™-2006 to include a subGHz PHY appropriate for use in Japan.

8.14 Study Group MBAN

A new area is being studied in the IEEE 802.15 WG. The Medical body area network (MBAN) is emerging as an area of interest for this standards development group. The use of "on body" as well as "in-body" wireless transmission and the related concerns of health issues, and so on are being debated.

The BAN SG has received a lot of interest and has developed a strong following. It is not clear how much the study group has liaised with other the IEEE WGs, that is, 11073 Medical Device Communications,* and so on and/or nonIEEE WGs. This area has been studied before by the IEEE 802.15 WG but not to this current level and it is likely that based on their current level of contributions that a new project for a BAN-based MAC/PHY specification(s) will be approved and developed.

8.15 ZigBee Alliance

The ZigBee alliance has published their specification as well they have created the prerequisite activities for compliance testing and certification; this has begun to enable the IEEE Std 802.15.4™-2006 wireless sensor networking technology. With a range of over 100 m, data rates of up to 250 Kbits/s and support for up to 65,000 nodes in star, cluster or mesh configurations, ZigBee targets low-data-rate, low-power applications ranging from home lighting control and automation to meter reading and industrial automation.

* http://www.ieee1073.org/standards/11073-20301/11073-20301.html

8.16 Alliance versus SDOs

There are various WPAN industry consortia that are working in concert with standards development organizations (SDO), for example, IEEE, and so forth to develop lower layer consensus-based WPAN standards and/or are developing industry specifications that reach beyond the charter of IEEE. Examples of these groups are Bluetooth SIG, ZigBee Alliance, and so on. Potentially emerging radio technologies such as Wibree from Nokia could be contributed to IEEE too.

However, some in the wireless industry consider that the traditional SDOs are not keeping pace with the rate of change and that industrial consortia offer a more controlled intellectual property rights (IPR) environment and shorter time to market.

The recent problems experienced by IEEE 802 in Project 802.15.3a and 802.20 are good examples of these problems, that is, SDO affiliation issues versus industrial sponsorship issues. The ability of the IEEE to manage individual voter behavior versus industrial sponsorship of individuals suggest that there may be some validity to these concerns.

The ability of an SDO to allow for the development of "open technical standards" is generally an accepted fact, that is, where all stakeholders may participate in the standards development process, all interests are discussed and agreement found, no domination, balloting and an appeals process may be used to find resolution, and holders of IPR must identify themselves during the standards development process. However, industry consortia, when they are well represented by the ecosystem, are able to closely emulate these "open" attributes and develop an acceptable "open technical specification." However, in industry consortia while all do not have to access committee documents, drafts prior to the IPR review; all have access to the completed specification.

These industry consortia, alliances, groups, and so on are proliferating and making it hard to keep up on the latest WPAN technologies. The following section tries to indicate a few ways for the reader to consider WPAN decision making although in the end it is likely that only by joining and participating one will learn how to take advantage of these exciting new technologies.

8.17 WPAN Decision Making

The process of choosing which wireless standard best addresses your application and/or system requirements remains complex and is based on a variety of issues, for example, mains versus battery, cost, platform integration, reuse of existing circuitry from multiple or an existing radio, average power consumption versus data rate, and so on.

Figure 8.6 describes the IEEE lower layer specifications available in their growing family of consensus-based standards. For system designers that are interested in wireless technologies this flow chart might help.

In addition, Figure 8.7 describes the various wireless networking technologies, for example, WPAN, WLAN the average power consumption versus data rate chart might be of interest as it tries to plot the existing as well as the emerging wireless networks in relative terms.

The areas that are in light grey (within dotted oval-shaped circle) are indicative of emerging WPANs or sensor network standards and specifications. These networks were overlayed onto the traditional domains for the sensor also known as RFID, WPANs, and WLANs areas.

8.18 Conclusion

Again, since the last printing of the handbook, we have seen an explosion of applications in the WPAN space. The dominant WPAN technology in use today is the Bluetooth wireless technology.

The Bluetooth SIG appears to be achieving their goal of growing their base and uniting the "other" wireless technologies under a Bluetooth umbrella. This is primarily being achieved by leveraging their successful vertical protocol stack approach, that is, profiles and their brand recognition in the computer networking, consumer electronics, automotive, medical, sensor, and so forth areas they enhance their

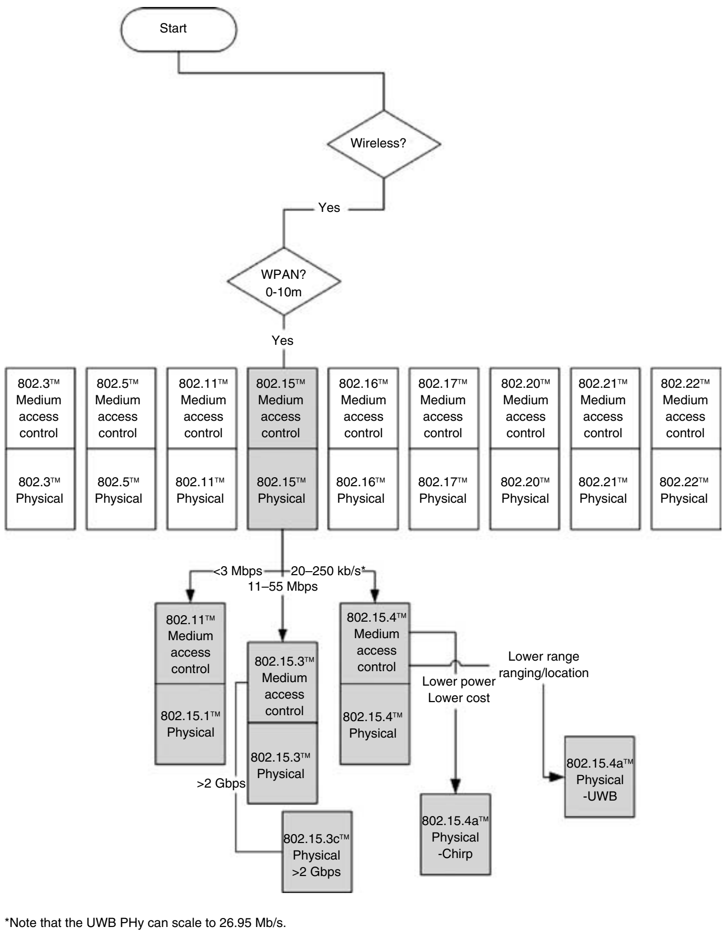


FIGURE 8.6 IEEE 802 family flow chart—WPAN.

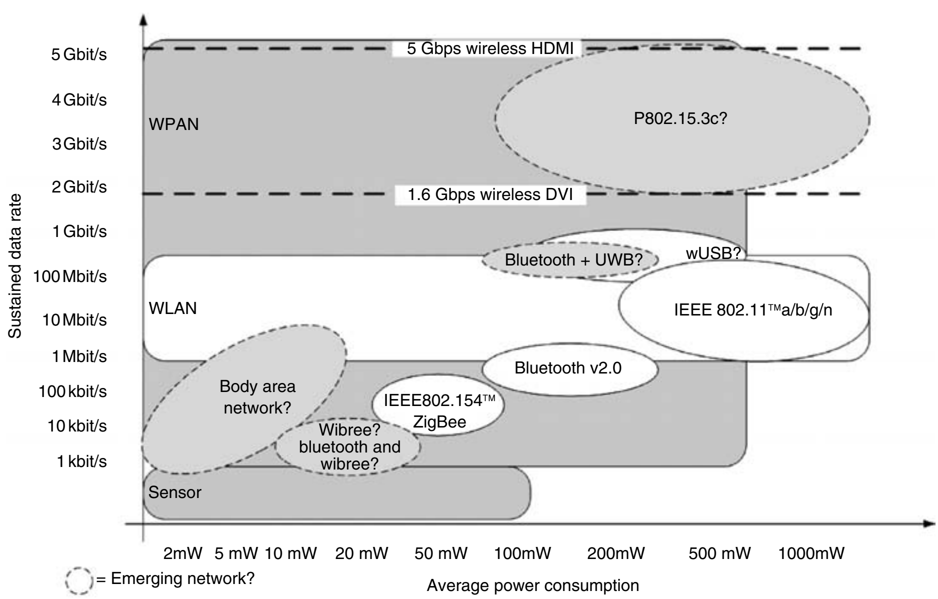


FIGURE 8.7 Average power consumption versus sustained data rate.

overall wireless experience. The Bluetooth RF (physical layer) currently operates at 2.4 GHz. The system employs a frequency hop transceiver to combat interference and fading and provides many FHSS carriers. RF operation uses a shaped, binary frequency modulation to minimize transceiver complexity. The symbol rate is 1 Ms/s supporting the bit rate of 1 Mbps or, with EDR, a gross air bit rate of 2 or 3 Mbps. These modes are known as Basic Rate and EDR, respectively. However, as an umbrella technology the Bluetooth + UWB will extend the frequency of interest to >6 GHz for the UWB transport and a gross air bit rate >3 Mbps, for example, 24, 50, 100 Mbps and beyond via other frequencies presumably this high speed transport will be controlled via the original Bluetooth wireless technology.

In summary, there is a shift in the epicenter of voice and data networking from the "local" to "personal"; our mobile devices (cell phone, PDA, PMP, digital camera) are becoming smaller and networked via multiple communications technologies. The role of the traditional PC is being relegated to fixed locations. The WPAN technologies are enabling technologies. Adopting these technologies allow us to accomplish more with less every day. It is an exciting technology and it appears that it will continue to be a significant and useful tool for ourselves as well as our children. The mobile devices are penetrating the population by morphing into fashion accessories, that is, color coordinated cell phone camera and headset, and so on. Finally, the explosive growth of the user-generated content or social networks of MySpace and YouTube are just the precursors for even more pervasive personal networks that will transform the world economies and at the center of this growth will be the WPAN.

9

Analog Fiber-Optic Links

William D. Jemison
Lafayette College

Arthur C. Paolella
Artisan Laboratories Corporation

9.1 Introduction to Analog Fiber-Optic Links

Analog fiber-optic links are used in many microwave-photonic applications. Microwave photonics is a field that uses light waves to aid in the transmission, signal generation, or signal processing of high-frequency analog signals. These signals may be in the radio frequency (RF), microwave, or millimeter-wave bands. The analog fiber-optic link is the fundamental building block for many microwave-photonic applications. A microwave analog fiber-optic link consists of an optical transmitter, a length of fiber, and a photodetector as shown in Figure 9.1.

Traditionally, the optical transmitter uses intensity modulation of a laser source to impress a microwave signal onto an optical carrier frequency. The modulated light is then transmitted over a low-loss fiber-optic cable and detected by a high-speed photodetector. The intensity modulation may be done by directly

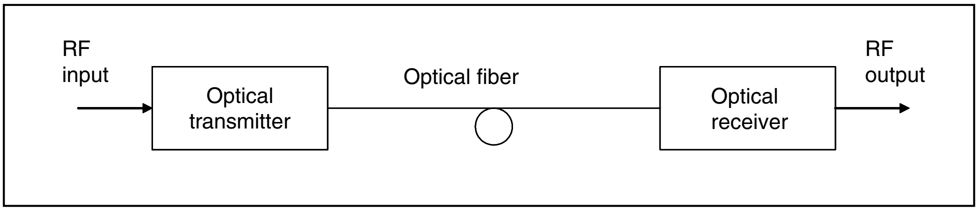


FIGURE 9.1 Block diagram of a generic analog fiber-optic link.

TABLE 9.1 Externally Modulated Link Parameters

Link Parameter	Value
P_{laser}	10 mW
K_{L}	0.5 dB
L	2 dB
K_{D}	0.5 dB
R	0.8 A/W
V_{π}	3 V
L_{M}	−6 dB
$R_{\text{S}}, R_{\text{M}}, R_{\text{L}}$	50 Ω

modulating the current of a semiconductor laser or by externally modulating a continuous-wave (CW) semiconductor or solid-state laser source with an RF signal. Common external modulators include the Mach-Zehnder modulator (MZM) and the electro-absorption modulator (EAM) (Table 9.1).

The implementation of analog microwave-photonic systems relies heavily on the use of optical components developed for digital fiber-optic telecommunications. For example, the wavelength used for early analog fiber-optic links was 850 nm, which is the wavelength emitted by GaAs semiconductor lasers that were used in the first generation of digital fiber-optic telecommunications. As the fiber-optic telecommunications industry migrated to 1300 nm and eventually to 1550 nm to take advantage of lower fiber dispersion and the development of the erbium doped fiber amplifier (EDFA), respectively, microwave photonics followed suit to take advantage of the performance and availability of the optical components developed at these longer wavelengths.

The advantages of analog fiber-optic links include immunity to electromagnetic interference, small size, light weight, and low loss in the optical fiber. An obvious application of these links is for "cable replacement," that is, the transmission of analog or digitally modulated RF, microwave, or millimeter-wave signals. Many existing and emerging commercial and military applications exist. For example, while the use of analog fiber optics is well established in the CATV industry, analog signal transmission over fiber is still not widely used in the telecommunications industry. However, services and technologies such as PCS, Broadband Wireless Internet, Digital Video, and Passive Optical Networks are creating new opportunities for fiber-optic codistribution of analog signal and digital data for long haul and metro access. Recently, a DWDM broadband photonic transport system was able to meet the requirements for both IS-95 Personal Communications Services operating at 1.9 GHz and Broadband Wireless Internet operating over the band of 2.5–2.7 GHz. Each DWDM channel operates from 1 to 3 GHz and transports services up to 80 km while reducing the complexity of the system architecture and base stations [1].

Analog fiber-optic links also have been used in antenna remoting applications and have been studied extensively for phased-array applications. New developments for the next generation of communication satellites and space based radar systems will employ active phased-array antennas; fiber-optic distribution has been proposed for the transmission of microwave signals from the satellite bus to the phased-array antenna. The advantages of using this technology are higher phase stability with temperature due to the fiber mechanical properties; lighter weight and flexibility of the fiber; and smaller size of the optical interconnect and optical power dividers. The use of fiber also allows for a high level of integration with antenna elements thus reducing size and weight [2]. Although these new developments show promise, the power consumption associated with fiber-optic distribution is a significant drawback compared with passive signal distribution. To make fiber-optic signal distribution competitive with other transmission technologies, high efficiency fiber-optic links are being developed that can significantly reduce the prime power consumption of the distribution system [3].

There also are applications of analog fiber-optic links in various RF testing applications. Radar and communication system tests typically require the use of outdoor ranges to determine system performance. Using an outdoor range is expensive and time consuming, so alternate solutions are desirable. Fiber-optic-based time delays can be used in target simulators for radar testing and calibration [4]. Analog fiber-optic links can also be used to test avionic and ground based systems in high electro magnetic interference (EMI) environments that require isolation between the platform under test and the data collection center. Analog fiber-optic links are also finding new applications in commercial and military short-haul communication systems. New commercial services such as fiber-to-the-premise for broadband to the business and home, broadband wireless Internet, and digital video services currently are being installed. The service providers and installers of these new systems require new methods of testing fiber-optic link performance over distances ranging from a few meters to 1 km. The optical time-domain reflectometers (OTDRs) traditionally used for testing long-haul fiber-optic communication systems have serious performance limitations when used to test these newer short-haul links due to dead-zones that can be as long as 10 m. New photonic test systems have been developed that have no dead-zone problems and have event dead-zones of less than 5 cm [5].

In addition to signal distribution and testing applications, variations of the analog fiber-optic link architecture can be used to realize photonic implementations of microwave circuits and subsystems [6]. These include microwave-photonic realizations of oscillators [7], filters [8], mixers [9], phase shifters [10], and time delays [11], as well as microwave-photonic implementations of phased-array manifolds [12], hybrid fiber/coax CATV distributions systems [13], and radio over fiber networks [14]. The advent of silicon photonics promises to drastically reduce the price of optical systems [15].

In summary, new applications continue to emerge in which analog fiber-optic link transmission can be employed. The primary goal of this chapter is to present the analog fiber-optic link in a manner that is familiar to the RF and microwave engineer such that he/she may evaluate analog fiber-optic link performance, determine the impact of link usage in RF, microwave, and millimeter-wave systems, and develop a greater awareness of the applications of microwave photonics.

9.2 Fiber-Optic Link Analysis Approach

A comprehensive treatment of optical components and fiber-optic link performance is beyond the scope of this chapter. We will limit the presentation to fiber-optic link architectures that use resistive impedance matching and contain neither optical nor RF amplification. The treatment of optical components is limited to a discussion of basic operation and the presentation of circuit models that aid in the development of the link modeling and the understanding of link performance. The reader is referred to two excellent texts devoted to a more comprehensive treatment of other link architectures and microwave-photonic components [16,17].

A general analysis approach for analog fiber-optic links is performed in three stages. First, we will develop expressions for the small-signal optical power emitted from the optical transmitter as a function of a high-frequency electrical excitation; the form of the excitation is laser current in the case of the directly modulated link and modulator voltage in the case of an externally modulated link. Next, we will account for optical effects including optical loss and fiber dispersion. Finally, we will develop expressions that convert the optical power delivered to a photodetector into an RF power delivered to the load.

This approach leads to the following "controlled source" model of a fiber-optic link shown in Figure 9.2. The optical transmitter is modulated by the RF signal source, v_s, to produce an optical output signal, p_{tx},

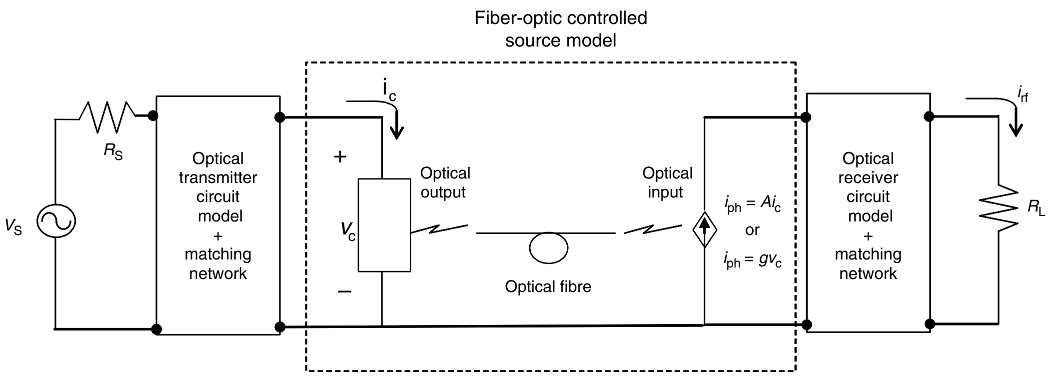


FIGURE 9.2 Controlled source circuit model for an analog fiber-optic link. The photocurrent may be current controlled or voltage-controlled depending on the type of link.

which is a function of an electrical current, i_c, or voltage, v_c. This optical power is coupled into an optical fiber where it is transmitted down the fiber and coupled into a high-speed photodetector. The optical power illuminating the photodetector, p_{rx}, produces a photocurrent, i_{ph}. A portion of this current is the resulting load current, i_{rf}, which is delivered to the load resistor, R_L.

Specifically, we will see that for the case of a directly modulated link, the controlled source is a current controlled current source with a link short-circuit current gain, A, which accounts for all electro-optical conversion, optical loss, and optical-electrical conversion. For the case of an externally modulated link, the controlled source is a voltage-controlled current source with a link transconductance, g, which accounts for all electro-optical conversion, optical loss, and optical-electrical conversion. The electric circuit elements in the "controlled source" model the electrical characteristics of optical transmitter and receiver.

In this manner, we can reduce the fiber-optic link to a simple electrical circuit that is easy to analyze, is familiar to the RF designer, and whose performance can be simulated in any commonly used RF simulator. Furthermore, the reader will recognize that this approach readily allows advanced circuit models for the optical transmitter and/or receiver to be incorporated into the simulation if better accuracy or the prediction of second-order effects is required.

9.2.1 Basic Analog Fiber-Optic Link Performance Specifications

The three primary analog fiber-optic link specifications are link gain, noise figure (NF), and dynamic range. As with RF amplifiers, there are several definitions of gain. We will use transducer gain, G_T, which is defined as the ratio of the RF power delivered to the load, p_{rf}, to the RF power available from the source, p_{avs}. The RF power available from the source is the power that is delivered to a load that is matched to the generator source impedance.

$$G_T = \frac{p_{rf}}{p_{avs}}.$$

The definition of link NF is analogous to that used for microwave devices. It is defined as the ratio of the signal-to-noise ratio at the input of the link, S_{in}/N_{in}, to the signal-to-noise ratio at the output of the link, S_{out}/N_{out}

$$\text{NF} = \frac{(S_{in}/N_{in})}{(S_{out}/N_{out})}$$

There are several definitions for dynamic range including compression dynamic range (CDR) and spurious-free dynamic range (SFDR). For externally modulated fiber-optic links employing the MZM, closed-form expressions may be derived for dynamic range assuming that the MZM limits the link linearity. For most other types of links, a series expansion of the transfer function can be used to predict the link dynamic range.

9.3 Analog Fiber-Optic Link Components—Semiconductor Lasers

Both solid-state and semiconductor lasers have been used in analog fiber-optic link designs. Solid-state lasers are often used for externally modulated fiber-optic links since their high output power and good noise characteristics lead to excellent link performance. Semiconductor lasers, on the other hand, are cheaper, more compact, and compatible with higher levels of integration than the solid-state laser. They can be either directly modulated making them usable for directly modulated fiber-optic links or operated CW for externally modulated fiber-optic links. Commercially available semiconductor lasers are pigtailed, or connected to a fiber-optic cable, using an automated alignment procedure to ensure good coupling efficiency into the fiber. Practically, the lasers must be temperature controlled as their wavelength and output power are sensitive to temperature.

9.3.1 Basic Laser Operation for Link Analysis

A simplified laser transmitter is shown in Figure 9.3 and a simplified optical power versus input current (L–I) curve is shown in Figure 9.4. Direct intensity modulation of the semiconductor laser may be accomplished by superimposing an RF current on the DC bias current to intensity modulate the optical carrier. The DC bias current, I_C, and RF signal, i_c, are injected into the laser via a bias tee producing an instantaneous laser current, $i_C = I_C + i_c$. This laser current produces an instantaneous optical output power, $p_{tx} = P_{TX} + p_{tx}$. The RF input can be ignored if the laser is to be operated in a CW mode for an externally modulated link. In this case, the application of a constant DC current, I_C, results in a constant optical output power, P_{TX}.

The semiconductor laser may be treated as a forward-biased diode that emits light once a threshold current, I_{th}, is exceeded. Therefore, the semiconductor laser is a device that converts injected electrons into emitted photons. The injected current, I_C, may be expressed as the product of the electron charge, q,

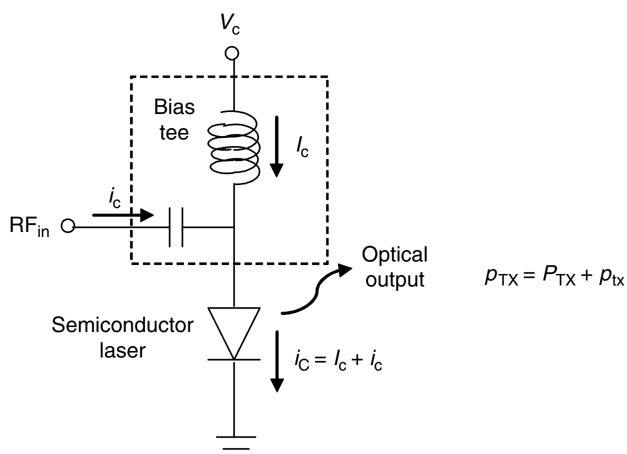


FIGURE 9.3 Simplified block diagram of a semiconductor laser transmitter. The RF input can be ignored if the laser is operated in a continuous-wave mode.

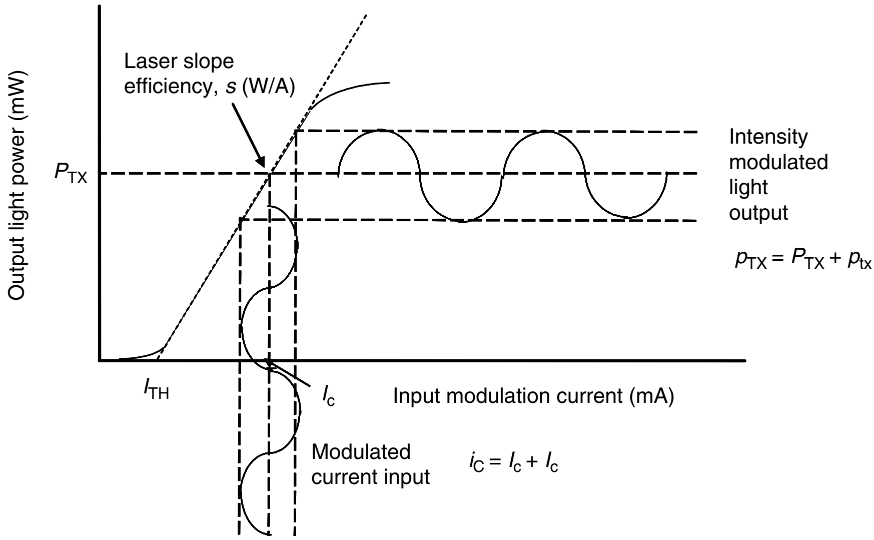


FIGURE 9.4 Ideal optical output power versus input current (P–I) transfer function for a semiconductor laser.

and the electron injection rate, Φ_e.

$$I_c = q\Phi_e$$

The photon energy, E_{ph}, is equal to the bandgap energy, E_g, of the laser material and is given by the following equation where h is Planck's constant, ν is the optical frequency, λ is the optical wavelength, and c is the speed of light.

$$E_{ph} = E_g = h\nu = \frac{hc}{\lambda_{optical}}$$

Therefore, the lasing wavelength is determined by the bandgap energy of the laser material. Historically, the wavelengths used for microwave photonics have tracked those used in commercial telecommunications [18]. Initial work started at 850 nm using GaAs lasers, moved to 1300 nm lasers, and is currently at 1550 nm using InGaAsP lasers [19].

The optical power, P_{TX}, emitted from the laser junction is equal to the product of the photon energy, E_{ph}, and the stimulated emission photon rate, Φ_{ph}.

$$P_{TX} = E_{ph}\Phi_{ph} = h\nu\Phi_{ph}$$

Therefore, we may define the laser internal quantum efficiency, η, as the ratio of emitted photons to injected electrons.

$$\eta = \frac{\Phi_{ph}}{\Phi_e} = \frac{(P_{TX}/h\nu)}{(I_c/q)}$$

The maximum internal quantum efficiency is 1 (i.e., one photon generated for each injected electron). The external quantum efficiency may be much less than the internal quantum efficiency due to the inherent difficulty of coupling the laser output, which has an elliptical beam profile, into a circularly symmetric single-mode fiber.

Ideally, the laser output power increases linearly with current above the lasing threshold, I_{TH}, and the ideal light–current (L–I) transfer function of the semiconductor laser is a piecewise linear function as shown in Figure 9.4. The slope of the P–I curve evaluated at the laser bias point is called the incremental laser slope efficiency, s, which has units of A/W and may be expressed as a function of the internal quantum efficiency and either the optical frequency or wavelength.

$$s = \left.\frac{dp_{\mathrm{TX}}}{di_{\mathrm{C}}}\right|_{I_{\mathrm{C}}} = \frac{\eta h\nu}{q} = \frac{\eta hc}{q\lambda_{\mathrm{opt}}}$$

9.3.2 Semiconductor Laser Models for Link Analysis

The use of a simple laser circuit model is sufficient for deriving the equations that describe basic analog fiber-optic link performance. The simplest model for the laser, as for any forward-biased diode, is a resistor in series with a DC voltage source where the resistor represents the on resistance and the voltage source represents the potential drop across the diode junction. For the laser, most of the laser voltage appears across the resistor, so to the first order, the laser may be modeled by single resistor, R_{i}. Assuming ideal RF and DC coupling into the laser (i.e., the bias tee capacitor is a perfect short and the inductor is an open circuit), the matching problem becomes the use of a simple series resistance, R_{M}, to match the RF generator source resistance, R_{S} (typically 50 Ω), to the laser resistance, R_{i} (typically several ohms), as shown in Figure 9.5.

The RF power delivered to the laser, p_{laser}, may then be easily computed as

$$p_{\mathrm{laser}} = i_{\mathrm{c}}^2 R_{\mathrm{i}} = \left(\frac{v_{\mathrm{s}}}{R_{\mathrm{s}} + R_{\mathrm{m}} + R_{\mathrm{i}}}\right)^2 R_{\mathrm{i}}$$

The power available from the source, P_{avs}, which is the power delivered to a matched load is

$$P_{\mathrm{avs}} = \frac{v_{\mathrm{s}}^2}{4R_s}$$

The simple laser model presented above is adequate for first-order fiber-optic link design and analysis. In an actual link design, it would be desirable to predict link frequency response, but the simple model is obviously inadequate as it contains no reactive elements. Furthermore, the simple laser model for link design does not predict optical resonance effects in the laser that lead to a peaking in the laser frequency response called relaxation oscillations. Numerous approaches have been developed to accurately model the performance of directly modulated semiconductor lasers by representing the laser rate equations with an equivalent circuit [20–25]. Equivalent circuit parameters may be extracted from measurements of the laser reflection coefficient, modulation response, and relative intensity noise (RIN). Most recently, an extraction procedure was proposed to extract both large and small-signal parameters based on closed-form equations, the measurement of the laser reflection coefficient, and the modulation response [26].

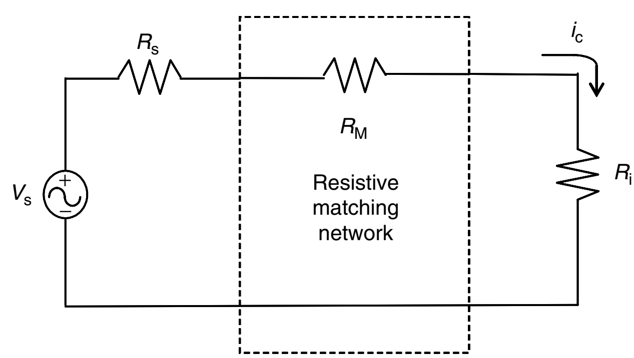


FIGURE 9.5 Ideal circuit model for a semiconductor laser assuming resistive matching.

9.4 Analog Fiber-Optic Link Components—Photodetectors

There are several photodetectors that are commonly used in photonics; however, the p-i-n photodiode is most commonly used for microwave photonics applications due to its good linearity and wide bandwidth.

9.4.1 Basic Photodetection

The purpose of the photodetector is to convert incident photons into photoelectrons that are collected to form a photocurrent. A photon has an energy that is proportional to the optical frequency.

$$E_{\text{ph}} = h\nu$$

The incident optical power is equal to the product of the photon energy and the incident photon rate, Φ_{ph}.

$$p_{\text{RX}} = E_{\text{ph}}\Phi_{\text{ph}} = h\nu\Phi_{\text{ph}}$$

Likewise, we can express the photocurrent as the product of the electron charge, q, and the generation rate of photoelectrons, Φ_{e}.

$$I_{\text{ph}} = q\Phi_{\text{e}}$$

Ideally, a photodetector converts each incident photon into a photoelectron that contributes to the photocurrent. This may occur when the photon energy is greater than the bandgap energy of the photodetector material. However, owing to reflection at the photodetector surface, surface recombination, and the spatial dependence of the photon absorption in the photodetector, not all incident photons produce a photoelectron. The ratio of the photoelectron generation rate, Φ_{e}, to the incident photon rate, Φ_{ph}, is called the detector quantum efficiency, η, which has a theoretical range from 0 to 1.

$$\eta = \frac{\Phi_{\text{e}}}{\Phi_{\text{ph}}} = \frac{(i_{\text{PH}}/q)}{(p_{\text{RX}}/h\nu)}$$

Here, q is the charge of an electron (1.6022×10^{-19} C), h is Plank's constant (6.626×10^{-34} J-s), and ν is the optical frequency. Therefore, we may express the detector photocurrent as

$$i_{\text{PH}} = q\Phi_{\text{e}} = q\eta\Phi_{\text{ph}} = \frac{q\eta}{h\nu}p_{\text{RX}} = rP_{\text{RX}}$$

We can see, therefore, the photocurrent is directly proportional to the incident optical power for an ideal photodetector. The constant of proportionality, r, is called the responsivity of the photodiode which has units of A/W. The responsivity is often expressed as

$$r = \frac{\eta q}{h\nu} \cong \frac{\eta\lambda}{1.24}$$

where λ is the optical wavelength in microns. The responsivity increases linearly with optical wavelength due to the photon energy dependence on wavelength. Since longer wavelength photons are less energetic than shorter wavelength photons, they produce less optical power for an equivalent incident photon rate.

Figure 9.6 shows an ideal photodiode I–L curve. Further, if a small RF intensity modulation is superimposed on the incident optical signal, it produces an RF output current superimposed on a DC bias. The DC bias may be removed from the RF signal via an appropriate RF coupling capacitor.

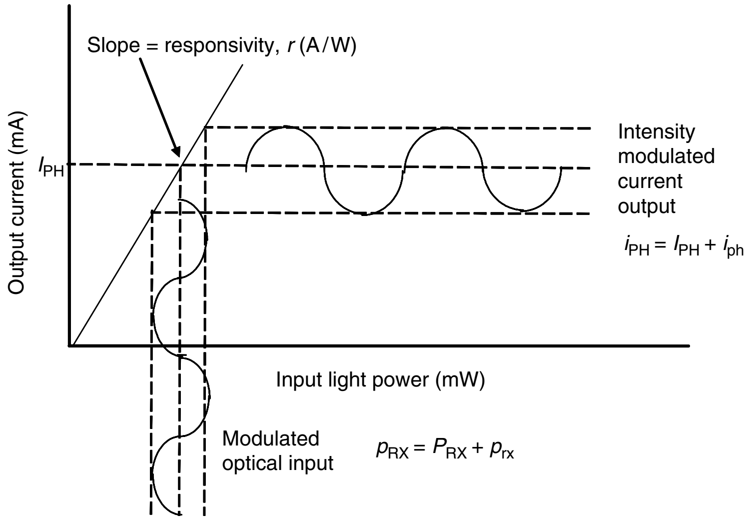


FIGURE 9.6 Ideal output current versus input light power (I–L) transfer function for a photodiode.

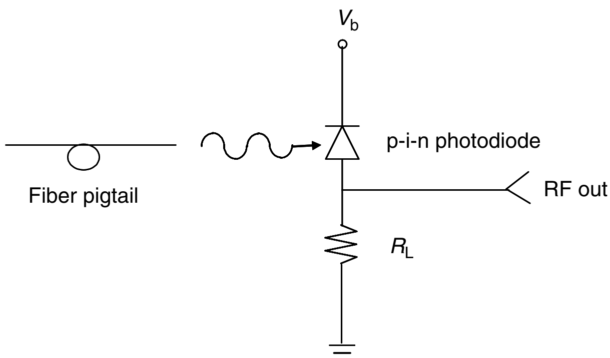


FIGURE 9.7 Block diagram of a photodiode.

It is important to note that the photodetector converts *optical power* to *electrical current.* If the electrical current is delivered to a resistive load, then the resulting *electrical power* is proportional to the *square of the optical power.* For microwave-photonic fiber-optic links that modulate the RF signal onto the optical intensity, 1 dB of optical power loss translates into 2 dB of RF electrical power loss.

9.4.2 Photodiode Models for Link Analysis

A typical p-i-n photodetector circuit is shown in Figure 9.7. Typically, the p-i-n photodiode is operated under several volts of reverse bias to ensure that photo-generated carriers reach their saturation velocity in the depletion region.

A simplified small-signal RF circuit model for the previous configuration can be obtained by replacing the p-i-n photodiode with an equivalent circuit model and setting the DC bias to RF ground. The resulting RF model, shown in Figure 9.8, consists of a current source, i_{ph}, which produces the photo-generated carriers, in parallel with a depletion resistance, R_j, and depletion capacitance, C_j, and a series resistance, R_s.

Practically, the diode electrical frequency response is determined by a trade-off between the transit time effects and RC effects. If the depletion region is made shorter, the photo-generated carries can be swept out of the diode more quickly. However, a shorter depletion region increases the photodiode capacitance. For the purposes of analog fiber optic-link analysis, we will connect the current source directly in parallel

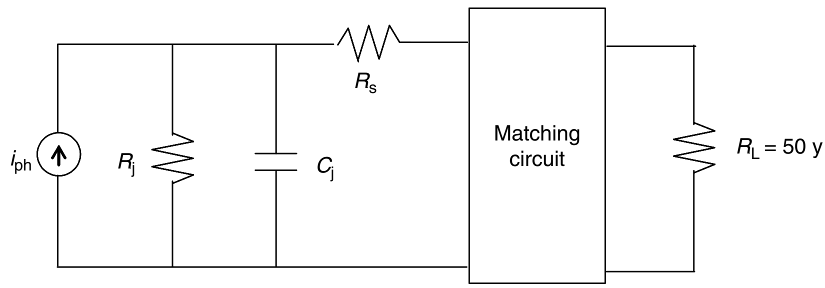


FIGURE 9.8 Simplified circuit model for a photodiode that can be used for analog fiber-optic link analysis.

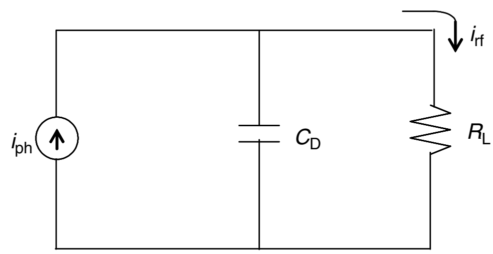


FIGURE 9.9 Simplified photodiode model for fiber-optic link analysis.

to an RC network that behaviorally defines the 3 dB frequency of the photodiode.

$$\omega_{3dB} \cong \frac{1}{\tau_D} = \frac{1}{R_L C_D}$$

The long wavelength cutoff of the detector is determined by the bandgap of the photodetector material (photons with wavelengths longer than the bandgap energy will not be absorbed). The short wavelength response of the detector degrades when photons are absorbed at the photodetector surface rather than in the depletion region.

This behavioral approach assumes that the depletion resistance is much greater than the load resistance and that the frequency response of the photodiode may be described by a single pole. Therefore, our simplified photodiode model for fiber-optic link analysis is given in Figure 9.9.

The RF load current may be easily computed as

$$i_{rf} = \frac{i_{ph}}{1 + s\tau_D}$$

Numerous high-speed photodetector designs have been realized both in the traditional vertically illuminated lateral configuration and the edge-coupled waveguide configuration and bandwidths in excess of 100 GHz have been reported [27]. In addition to bandwidth, the photodetector linearity is also important [28]. Trends in high-speed photodetectors include improving the bandwidth-efficiency product and obtaining high saturation current. Recently, a detector has been developed that can handle very high incident modulated optical power and it produced 20 dBm of RF output power directly from the detector [29].

9.5 Analog Fiber-Optic Link Components—External Modulators

The MZM and the electro-absorption (EA) modulator are the most commonly used devices for externally modulated analog fiber-optic links. The MZM works on the principle of optical interference while the EA modulator works on the absorption of light via a bulk semiconductor effect using the Franz Keldysh effect or via band edge modulation using the quantum confined Stark effect (QCSE).

9.5.1 Mach-Zehnder Modulator

There are several basic variants of the MZM including low-frequency lumped element modulators and high-frequency traveling modulators. There is also a closely related interferometric modulator that uses an optical directional coupler to combine the optical signals [30]. The fundamental operation of each of these devices is based on dynamically controlling the optical phase in one arm of an optical interferometer using the linear electrooptic effect, or Pockel's effect. This effect allows the conversion of optical phase modulation into optical intensity modulation. Modulator materials include lithium niobate ($LiNbO_3$) [31], polymers [32], and semiconductors [33]. The RF electrode design is an important consideration in the MZM design. For low-speed devices, the RF electrode can be considered as a lumped element capacitor. However, for higher-speed designs, the velocity of the optical wave must be matched with the velocity of the RF signal [34]. A generic MZM configuration is shown in Figure 9.10.

The optical input signal is split into two paths. A DC voltage is applied to the upper arm of the MZM and an RF signal is applied to the lower arm. Each voltage will induce its own electric field, E, that will change the index of refraction, $n(E)$, in the region of the optical waveguide where the field is present.

$$n(E) = n - \frac{1}{2} r n^3 E$$

Here, r is the linear electrooptic coefficient. The index change, in turn, produces a phase shift, $\Delta\phi(E)$, in the optical signal as it propagates a distance l in the device

$$\Delta\phi(E) = \phi_0 - \frac{2\pi}{\lambda_0} n(E) l$$

where ϕ_0 is the phase shift that would occur in the absence of the applied field and λ_0 is the free space optical wavelength. Alternatively, we may express the phase shift in terms of the applied voltage and a quantity known as the half-wave voltage, V_π.

$$\Delta\phi(V) = \phi_0 - \pi \frac{V}{V_\pi}$$

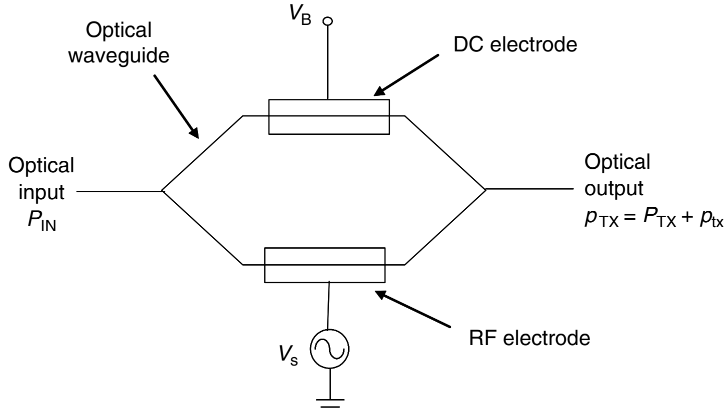


FIGURE 9.10 Basic Mach-Zehnder Modulator (MZM).

V_π is the voltage that results in an optical phase shift of 180° in one arm of the MZM causing a change in the MZM output from a condition of constructive interference (maximum light intensity) to a condition of destructive interference (minimum light intensity). Assuming that the electric field is applied uniformly across an optical waveguide dimension of d, we may express V_π in terms of the physical parameters of the MZM by recognizing that $V = Ed$ for a uniform field across the optical waveguide.

$$V_\pi = \frac{d}{l}\frac{\lambda_0}{rn^3}$$

Intuitively, a lower V_π is desirable as less RF voltage is required to produce the desired interferometric effect in the MZM. As will be seen later, a lower V_π is also desired to increase the gain of externally modulated fiber-optic links. Thus, for a given optical wavelength, we can see that V_π may be reduced by increasing the optical/RF interaction length, l, using a material with a higher index of refraction, n, or a higher electrooptic coefficient, r. However, as mentioned previously, the velocity of propagation of the optical and RF signals in the device much be matched with the region where they interact in order for the MZM to maintain good bandwidth performance. The velocity of propagation of the optical signal is determined primarily by the index of refraction of the MZM material. The velocity of propagation of the RF signal is determined by the effective dielectric constant of the MZM material and the electrode structure.

9.5.1.1 MZM Transfer Function

Assuming that there is an equal power split in the MZM and ignoring any splitter and waveguide losses for the time being, the intensity transfer function of the interferometer is given by

$$I_0(V) = I_i L_M \cos^2\left(\frac{\varphi(V)}{2}\right) = I_i L_M(1 + \cos(\varphi(V)))$$

where L_M accounts for the modulator optical insertion loss. The transfer function may be expressed in terms of a fixed optical phase shift, ϕ_o, between the MZM arms and the half-wave voltage, V_π.

$$I_0(V) = I_i L_M\left(1 + \cos\left(\phi_0 - \pi\frac{V}{V_\pi}\right)\right)$$

If the optical path lengths are designed such that there is a 90° difference in optical phase between the arms of the interferometer (i.e., $\phi_0 = 90°$), then the transfer function may be written as

$$I_0(V) = I_i L_M \cos^2\left(\frac{\varphi}{2}\right) = I_i L_M\left(1 + \sin\left(\pi\frac{V}{V_\pi}\right)\right)$$

For analog applications, the modulator is typically biased at quadrature, meaning that the RF signal is applied to the modulator input on the linear portion of the transfer function as shown in Figure 9.11. However, a low-biasing scheme has also been proposed, which can increase dynamic range performance for sub-octave links [35]. Note that there are two regions in the MZM transfer function that have a linear slope. The positive slope results in a noninverting modulation, while the negative slope results in a 180° phase shift. This has been exploited for RF vector modulation [36] and optical MZM linearization. The closed-form expression of the MZM modulator transfer function will allow the derivation of a closed-form expression for link linearity when an MZM is used.

Expressing the applied voltage as the sum of a DC bias, V_B, and RF signal, V_c, we get

$$p_{TX} = L_M\frac{P_{IN}}{2}\left(1 + \sin\left(\frac{\pi\,(V_B + V_c)}{V_\pi}\right)\right)$$

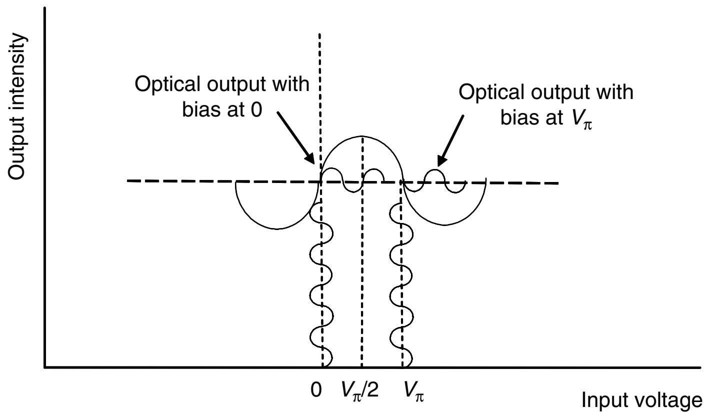


FIGURE 9.11 Ideal MZM intensity transfer function.

The sinusoidal term may be expanded such that

$$p_{\mathrm{TX}} = L_{\mathrm{M}} \frac{P_{\mathrm{IN}}}{2} \left[1 + \sin\left(\frac{\pi V_{\mathrm{B}}}{V_\pi}\right) \cos\left(\frac{\pi v_{\mathrm{c}}}{V_\pi}\right) + \cos\left(\frac{\pi V_{\mathrm{B}}}{V_\pi}\right) \sin\left(\frac{\pi v_{\mathrm{c}}}{V_\pi}\right) \right]$$

Under a small-signal assumption

$$\cos\left(\frac{\pi v_{\mathrm{c}}}{V_\pi}\right) \cong 1$$

$$\sin\left(\frac{\pi v_{\mathrm{c}}}{V_\pi}\right) \cong \frac{\pi v_{\mathrm{c}}}{V_\pi}$$

such that

$$p_{\mathrm{TX}} = L_{\mathrm{M}} \frac{P_{\mathrm{IN}}}{2} \left[\left(1 + \sin\left(\frac{\pi V_{\mathrm{B}}}{V_\pi}\right) \right) + \left(\cos\left(\frac{\pi V_{\mathrm{B}}}{V_\pi}\right) \frac{\pi v_{\mathrm{c}}}{V_\pi} \right) \right]$$

The small-signal optical output power is, therefore

$$p_{\mathrm{tx}} = L_{\mathrm{M}} \frac{P_{\mathrm{IN}} \pi}{2 V_\pi} \cos\left(\frac{\pi V_{\mathrm{B}}}{V_\pi}\right) v_{\mathrm{c}}$$

9.5.1.2 MZM Circuit Model

For the purposes of basic link analysis, we can use a simplified traveling-wave modulator model for the MZM as shown in Figure 9.12.

In this model, it is assumed that the transmission line impedance, Z_0, load impedance, R_{L}, and source resistance, R_{S}, are all identical (typically 50 Ω). Thus, the input impedance, Z_{in}, looking into the modulator transmission line is R_{s}. Since the load is matched, there are no reflections and the incident wave voltage, V_0^+, is equal to one-half of the generator voltage, V_{s}. This will be the voltage that interacts with the electrooptic material.

$$v_{\mathrm{c}} = v_0^+ = \frac{v_{\mathrm{s}}}{2}$$

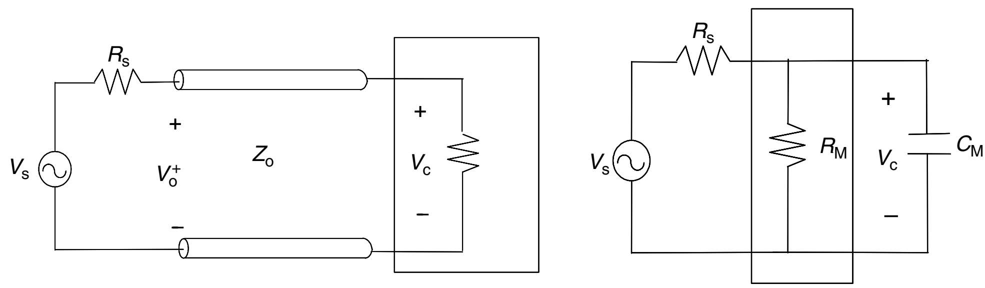


FIGURE 9.12 Simplified circuit model for the MZM modulator that can be used for analog fiber-optic link analysis. The model on the left is a traveling-wave model. The model on the right is a lumped element model but can also be used to behaviorally model a traveling-wave device.

We may "behaviorally" model the modulator frequency response by defining a simple single-pole low-pass frequency response to the modulator where $\omega_{3\text{dB}}$ is the 3 dB frequency of the MZM

$$v_c = v_0^+ = \frac{v_s}{2}\frac{1}{1+s\tau_M}$$

A lumped element MZM model, shown in Figure 9.12, uses a parallel resistance, R_M to match the MZM capacitance, C_M. In both the traveling-wave behavioral model and the resistively matched lumped element model the 3 dB frequency is

$$\omega_{3\text{dB}} = \frac{1}{\tau_M} = \frac{1}{R_M C_M}$$

The RF power delivered to the modulator can be expressed as follows where Z_P is the parallel load impedance.

$$p_{\text{rf}} = \frac{v_c^2}{Z_P} = \frac{v_s^2}{2R_M}\left(\frac{1}{1+s\tau_M}\right)^2$$

For $\omega \ll 1/\tau_M$ then we have

$$p_{\text{rf}} \approx \frac{v_s^2}{2R_M}$$

9.5.1.2.1 Linearization

Since the transfer function of the modulator is inherently nonlinear, the modulation depth must be kept small to ensure linearity. There have been many approaches to linearizing the MZM including electronic predistortion [37], electro-optical predistortion [38], and various optical linearization techniques including the use of series MZMs [39], parallel MZMs [40], modified directional couplers [41], polarization mixing [42], and the use of dual wavelengths [43].

9.5.1.2.2 Electroabsorption Modulator

The electroabsorption modulator (EAM) operates based on the QCSE where an applied voltage is used to shift the band edge that is near the long wavelength cutoff of the device [44,45]. A basic EAM configuration is shown in Figure 9.13.

The optical output power may be expressed as

$$p_{\text{TX}} = p_{\text{IN}} e^{-\Gamma L \alpha(v_C)} = p_{\text{IN}} e^{-\Gamma L \alpha(V_C)} e^{-\Gamma L \Delta\alpha(v_c)}$$

where Γ is a confinement factor that describes the fraction of incident light that is in the absorbing region of the EAM, L is the interaction length, and $\alpha(v_c)$ is the voltage-dependent absorption coefficient.

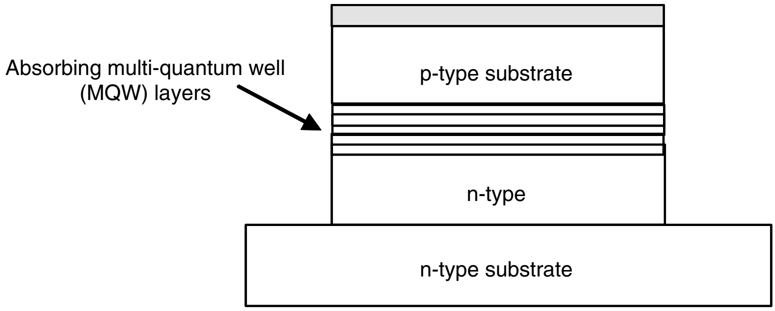


FIGURE 9.13 Basic electroabsorption modulator.

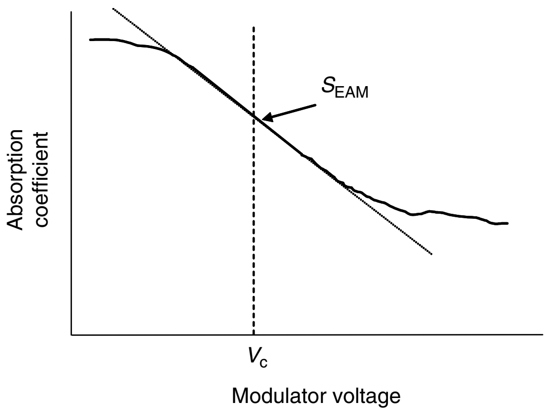


FIGURE 9.14 Insertion loss versus input voltage transfer function.

The applied voltage consists of both DC and RF components (i.e., $v_C = V_C + v_c$) where V_C is the DC bias and v_c is the RF voltage. The application of the DC bias introduces a fixed attenuation or loss

$$L_b = e^{-\Gamma L\alpha(V_b)}$$

Assuming small-signal operation about the bias point, we may compute the derivative of the absorption coefficient at the EAM bias point, V_C, as shown in Figure 9.14. A bias voltage of a couple of volts gives an insertion loss on the order of 25 dB.

$$\left.\frac{d\alpha}{dv_C}\right|_{V_C} = s_{eam}$$

Under this small-signal assumption, the change in the absorption coefficient, $\Delta\alpha$, is directly proportional to the applied RF voltage, v_c.

$$\Delta\alpha = s_{eam} v_c$$

Therefore, the EAM output power becomes

$$p_{TX} = p_{IN} L_b e^{-\Gamma L\Delta\alpha(v_c)} = p_{IN} L_b e^{-\Gamma L s_{eam} v_c}$$

Furthermore, for small-signal modulation, the exponential term containing the RF signal may be approximated by a series expansion (i.e., $e^x \cong 1 + x$) such that the EAM optical output power may be approximated as

$$p_{\mathrm{TX}} = p_{\mathrm{IN}} \mathrm{e}^{-\Gamma L \alpha(v_{\mathrm{C}})} = p_{\mathrm{IN}} L_{\mathrm{b}} \left(1 - \Gamma L s_{\mathrm{eam}} v_{\mathrm{c}}\right)$$

Furthermore, we may define a parameter, V_{α}:

$$V_{\alpha} = \frac{-1}{\Gamma L s_{\mathrm{eam}}}$$

Such that the small-signal EAM optical power is

$$p_{\mathrm{tx}} = p_{\mathrm{IN}} L_{\mathrm{b}} \frac{v_{\mathrm{c}}}{V_{\alpha}}$$

Note that when the applied RF voltage, vc_{m}, is equal to V_{α}, the EAM output optical power is twice that under its DC bias (i.e., with no RF voltage applied) condition.

9.5.1.3 Circuit Model

An equivalent circuit model may be used to predict the EAM performance [46]. For our purposes, we will assume a simplified model consisting of a parallel $R_{\mathrm{M}} C_{\mathrm{M}}$ network as shown in Figure 9.12. The shunt resistance of the EAM will be much higher than the typical source resistance of 50 Ω, so we may apply a simple matching network consisting of a parallel resistor, $R_{\mathrm{M}} = R_{\mathrm{S}}$. Thus, the simple circuit model of the EAM is identical in form to that of our simple MZM model. As we will see later, this will mean that the circuit analysis we will use to determine the basic link gain and NF will be identical for externally modulated links using the MZM and the EAM. Although the general analysis is the same for the MZM and EAM, there are limitations on the link gain and NF that are device specific [47]. Linearization techniques have also been employed for multi-quantum well EAM modulators [48]. EAMs may also be integrated with lasers on the same device to produce an electroabsorption-modulated laser (EML) [49].

9.6 Analog Fiber-Optic Link Gain Analysis

This section presents the analysis for analog fiber-optic link gain. Recall that the link gain will be specified using the transducer gain, G_{T}, which is defined as the ratio of the RF power delivered to the load, p_{rf}, to the RF power available from the source, p_{avs}.

$$G_{\mathrm{T}} = \frac{p_{\mathrm{rf}}}{p_{\mathrm{avs}}}$$

The RF power available from the source is the power that is delivered to a load that is matched to the generator source impedance as shown in Figure 9.15.

$$p_{\mathrm{avs}} = \frac{v_{\mathrm{s}}^2}{4R_{\mathrm{s}}}$$

9.6.1 Directly Modulated Link

For the directly modulated analog fiber-optic link we may employ the controlled source model shown in Figure 9.16. As described in previous sections, we will assume simple resistive series matching to the laser and ignore laser parasitics. We will also assume that the output resistance of the photodiode is much

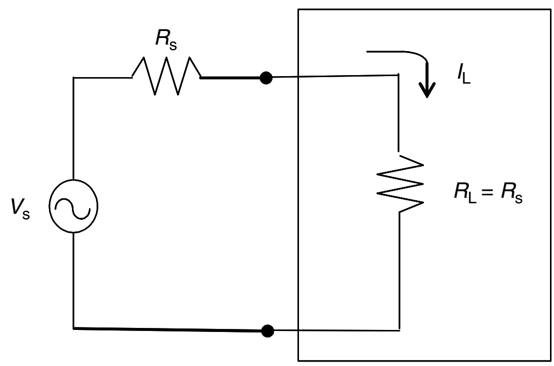


FIGURE 9.15 Circuit diagram for analysis of power available from the source.

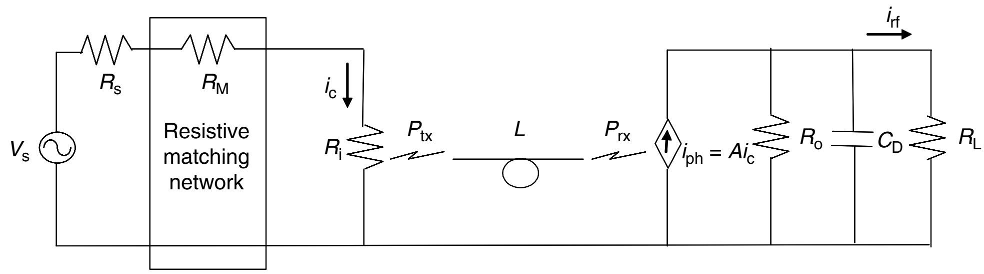


FIGURE 9.16 Current controlled current source model for directly modulated analog fiber-optic link analysis.

greater than the load resistance and that the detector shunt capacitance is selected to define a single-pole frequency response of the photodetector where

$$\omega_0 = \frac{1}{R_L C_D}$$

Using this circuit model, the small-signal laser current, i_s, may be computed as

$$i_c = \frac{v_s}{R_S + R_M + R_i}$$

If the series matching resistor, R_M, is selected to impedance match the source to the laser (i.e., $R_S = R_M + R_i$) then we have

$$i_c = \frac{v_s}{2(R_M + R_i)}$$

The optical power emitted by the laser, p_{tx}, is the product of the laser slope efficiency, s, and the RF laser current, i_c.

$$p_{tx} = i_c s$$

The optical power incident on the detector, p_{rx}, is the optical power emitted by the laser less the coupling loss from the laser into fiber, K_L, the fiber attenuation, L_F, and the coupling loss from the fiber into the photodetector, K_D.

$$p_{rx} = K_L L_F K_D p_{tx} = sK_L L_F K_D i_c$$

The photocurrent produced in the photodiode is equal to the product of the optical power incident on the detector and photodetector responsivity, r.

$$i_{\mathrm{ph}} = p_{\mathrm{rx}} r = (sK_{\mathrm{L}} L_{\mathrm{F}} K_{\mathrm{D}} r)\, i_{\mathrm{c}} = A i_{\mathrm{c}}$$

Thus we may define a short-circuit current gain, A, that is used in the controlled source and includes the optical to electrical conversion of the laser (i.e., the slope efficiency), all optical losses (i.e., coupling, fiber loss, and dispersion), and the optical to electrical conversion (i.e., the photodetector responsivity). It may also include a term to model the laser relaxation oscillation frequency and fiber dispersion if desired.

$$A = sK_{\mathrm{L}} L_{\mathrm{F}} D K_{\mathrm{D}} r$$

Likewise, the RF power delivered to the load may be computed as

$$p_{\mathrm{rf}} = i_{\mathrm{rf}}^2 R_{\mathrm{L}}$$

where the load current, i_{l}, may be computed via current division

$$i_{\mathrm{rf}} = \frac{A i_{\mathrm{c}}}{1 + j\omega R_{\mathrm{L}} C_{\mathrm{D}}}$$

The computation of the directly modulated link transducer gain, G_{TD}, is then easily computed as

$$G_{\mathrm{TD}} = \frac{A^2 R_{\mathrm{L}}}{(R_{\mathrm{M}} + R_{\mathrm{i}})(1 + j\omega R_{\mathrm{L}} C_{\mathrm{D}})^2}$$

For frequencies well below both the 3 dB frequency of the photodetector (i.e., $\omega \ll (R_{\mathrm{L}} C_{\mathrm{D}})^{-1}$) and the relaxation oscillation frequency of the laser we have

$$G_{\mathrm{TD}} = \frac{A^2 R_{\mathrm{L}}}{(R_{\mathrm{M}} + R_{\mathrm{i}})}$$

An examination of this expression shows several important points. First, the directly modulated link gain is independent of the average laser power. Thus, the use of a higher power laser will not improve the link gain. Second, all of the optical loss terms (i.e., coupling losses K_{D} and K_{L} and fiber loss L_{F}) are squared. The impact of this is that 1 dB of optical power loss in the link produces 2 dB of RF power loss. Finally, the use of resistive matching negatively impacts the gain of the link since R_{M} appears in the denominator of the link gain expression. Cox examines that case where the resistive matching is replaced by other types of matching networks including lossless transformers [16]. In these cases the gain can be improved, but the link bandwidth is reduced.

9.6.2 Externally Modulated Link Gain—MZM

For the externally modulated analog fiber-optic link we may employ the controlled source model shown in Figure 9.17. We will use the simple single-pole model for the MZM modulator described previously. We will also assume that the output resistance of the photodiode is much greater than the load resistance and that the detector shunt capacitance is selected to define a single-pole frequency response of the photodetector where

$$\omega_{\mathrm{D3dB}} = \frac{1}{R_{\mathrm{L}} C_{\mathrm{D}}}$$

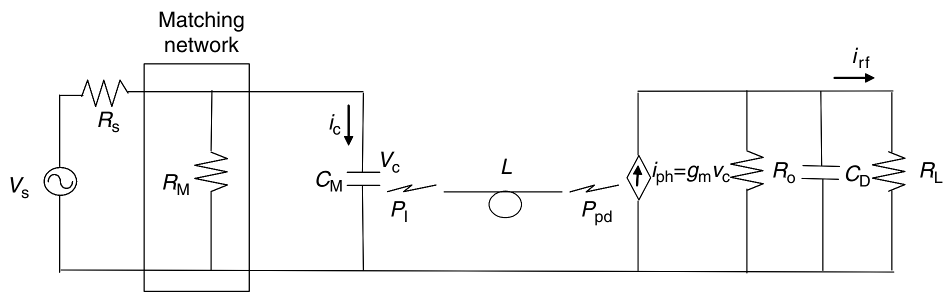


FIGURE 9.17 Controlled source model for link gain analysis of the externally modulated fiber-optic link. This topology is applicable to links employing either MZMs or EAMs.

The modulator voltage, v_M, can be determined by analyzing the input side of the controlled source model.

$$v_M = \frac{v_s}{2} \frac{1}{(1 + s\tau_M)}$$

for $R_S = R_M$ and where τ_M is related to the modulator model and 3 dB modulator frequency by

$$\omega_{M3dB} = \frac{1}{\tau_M} = \frac{1}{R_M C_M}$$

The optical power output of the MZM was derived previously and can be written as

$$p_{tx} = L_M \frac{P_{IN}\pi}{2V_\pi} \cos\left(\frac{\pi V_B}{V_\pi}\right) v_c$$

This power is reduced by the coupling loss from the modulator into the fiber, K_T, the fiber loss, L_F, and the coupling into the photodetector, K_D, to give optical power incident on the photodetector, p_{det}.

$$p_{rx} = p_{tx} K_M L_F K_D = (K_M L_F K_D)\, L_M \frac{P_{IN}\pi}{2V_\pi} \cos\left(\frac{\pi V_B}{V_\pi}\right) v_c$$

The photocurrent produced by optical power incident on the detector may then be computed as

$$i_{ph} = p_{rx} r = p_{tx} K_M L_F K_D r = (K_M L_F K_D r) L_M \frac{P_{IN}\pi}{2V_\pi} \cos\left(\frac{\pi V_B}{V_\pi}\right) v_c = g_{mzm} v_c$$

The transconductance of the controlled source model, g_m, that is used in the controlled source includes the electrical to optical conversion of the MZM, all optical losses (i.e., coupling, fiber loss, and dispersion), and the optical to electrical conversion (i.e., the photodetector responsivity).

$$g_{mzm} = (K_M L_F K_D r)\, L_M \frac{P_{IN}\pi}{2V_\pi} \cos\left(\frac{\pi V_B}{V_\pi}\right)$$

Likewise, the RF power delivered to the load may be computed as

$$p_{lrf} = i_{rf}^2 R_L$$

where the load current, i_l, may be computed via current division

$$i_{rf} = \frac{g_{mzm} v_c}{1 + j\omega R_L C_D}$$

The directly modulated link transducer gain, G_{mzm}, is then easily computed as

$$G_{mzm} = \frac{g_{mzm}^2 R_M R_L}{(1 + s\tau_M)^2 (1 + s\tau_D)^2}$$

For frequencies that are much less than the 3 dB frequencies of the modulator and detector, the link gain is

$$G_{mzm} = g_{mzm}^2 R_M R_L$$

The externally modulated link gain is proportional to the square of the average laser power. The use of a higher power laser will dramatically improve the link gain. Second, as in the case with the directly modulated link, all of the optical loss terms (i.e., coupling losses K_D and K_L and fiber loss L_F) are squared such that 1 dB of optical power loss result in 2 dB of RF power loss.

9.6.3 Externally Modulated Link Gain—EAM

Since the basic circuit model developed for the EAM is identical to that developed for the simple MZM, the link model for an externally modulated link employing EAM is identical to that using an MZM. Thus, the basic analysis is the same and we need to substitute only in the appropriate electro-optical transfer function to get the appropriate transconductance and hence link gain for the EAM case.

$$G_{eam} = \frac{g_{eam}^2 R_M R_L}{(1 + s\tau_M)^2 \, (1 + s\tau_D)^2}$$

$$g_{eam} = \frac{K_M L_f K_D r P_{laser} L_b}{V\alpha}$$

An examination of these expressions shows several important points, most of them very similar to the observations made for the MZM modulated link. The EAM may also potentially be integrated with the laser [50].

9.6.3.1 Dispersion-Induced RF Gain Penalty

Dispersion can cause a significant RF power penalty at the link output. It can affect either long-haul analog links operating in the microwave band or short-haul analog links operating in the millimeter-wave band [51]. Dispersion in single-mode fiber is the result of a combination of material dispersion and waveguide dispersion. The combination of these effects results in a propagation constant, β, that is a function of the optical frequency. This may be expanded into a Taylor series as shown below.

$$\beta(\omega) = \beta(\omega_c) + (\omega - \omega_c) \left.\frac{d\beta}{d\omega}\right|_{\omega_c} + (\omega - \omega_c)^2 \left.\frac{d^2\beta}{d\omega^2}\right|_{\omega_c} + (\omega - \omega_c)^3 \left.\frac{d^3\beta}{d\omega^3}\right|_{\omega_c} + \cdots$$

The group velocity is defined as

$$v_g = \left(\frac{d\beta}{d\omega}\right)^{-1}$$

The dispersion parameter is defined as

$$D = \frac{\mathrm{d}}{\mathrm{d}\lambda}\left(\frac{1}{v_\mathrm{g}}\right) = -\frac{2\pi}{\lambda^2}\left(\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right) = -\frac{2\pi f^2}{c}\left(\frac{\mathrm{d}^2\beta}{\mathrm{d}\omega^2}\right)$$

which has units of ps/(km-nm). A typical value of D is 19 ps/(km-nm) for SMF-28 single-mode fiber. An analysis of an intensity-modulated signal propagating over a dispersive fiber shows that the fundamental frequency at the photodetector output, p_rf, is proportional to [52]

$$p_\mathrm{rf} \propto \cos^2\left(\frac{\pi\lambda_\mathrm{c}^2 f^2 LD}{c}\right) = L_\mathrm{D}^2$$

where λ_c is the optical carrier wavelength, f is the RF frequency, L is the length of the fiber, and D is the dispersion parameter. Therefore, in order to incorporate a dispersion-induced RF power penalty in the link gain expression, we simply need to add L_D into the short-circuit current gain, A, for the directly modulated link, or the transconductance, g_mzm or g_eam, for the externally modulated links. For example, the short-circuit current gain for a directly modulated link incorporating a dispersion induced RF power penalty would be

$$A = sK_\mathrm{L}L_\mathrm{F}L_\mathrm{D}DK_\mathrm{D}r$$

Figure 9.18 shows a simulated dispersion induced power penalty as a function of modulation frequency for a 1550 nm externally modulated 79.6 km link with $D = 17$ ps/(km-nm) [53]. The plot shows that a significant dispersion penalty will occur for modulation frequencies above a few GHz.

It should be noted that the dispersion-induced RF power penalty described above should only be used for double sideband intensity modulation when the optical wavelength is not at the zero-dispersion wavelength. Typical examples would be traditional directly or externally modulated links operating at 1550 nm. Dispersive effects can be reduced via single-sideband techniques [54], by operating at the zero dispersion wavelength (~1310 nm for SMF-28 fiber), or by using either dispersion shifted fiber or a combination of regular and dispersion compensated fiber [55]. Polarization mode dispersion may also be problematic, particularly for millimeter-wave links [56].

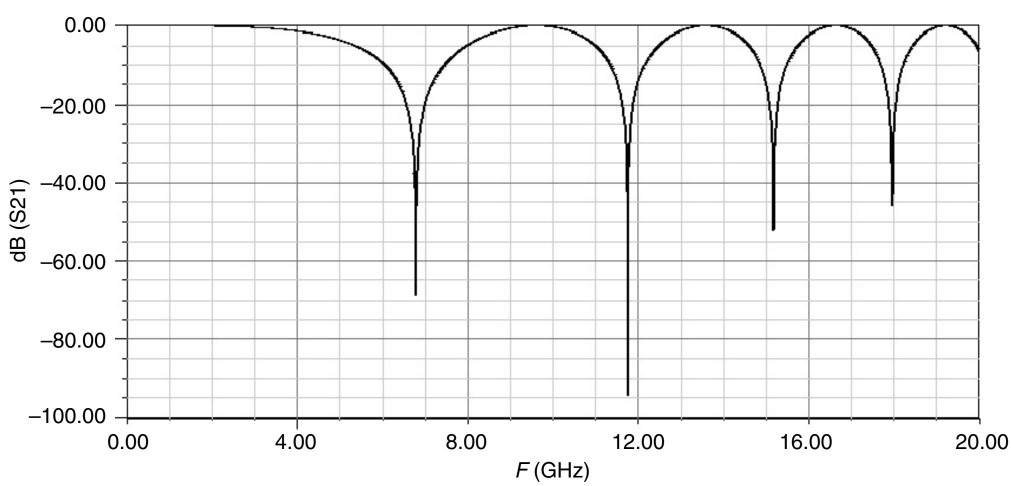


FIGURE 9.18 Simulated dispersion penalty for 79.6 km of single-mode fiber with $D = 17$ ps/(km-nm).

9.7 Analog Fiber-Optic Link Noise Figure Analysis

The definition of link NF is identical to that used for microwave devices. It is defined as the ratio of the signal-to-noise ratio at the input of the link, S_{in}/N_{in}, to the signal-to-noise ratio at the output of the link, S_{out}/N_{out}.

$$\mathrm{NF} = \frac{(S_{in}/N_{in})}{(S_{out}/N_{out})} = \left(\frac{S_{in}}{S_{out}}\right)\left(\frac{N_{out}}{N_{in}}\right) = \left(\frac{1}{G}\right)\left(\frac{N_{in}+N_{add}}{N_{in}}\right) = 1 + \frac{N_{add}}{GN_{in}}$$

The input noise, N_{in}, is the available noise power from RF input source due to the source resistor at 290 K. This can be easily computed as $N_{in} = kT_0B$ where k is Boltzman's constant, T_0 is the standard temperature of 290 K, and B is the bandwidth. Further, G is the link gain and N_{add} is the added noise power. Since the link gain was determined previously, the remaining challenge is to compute the added noise contributions due to thermal noise, shot noise, and RIN.

9.7.1 Thermal Noise

Thermal or Johnson noise arises due to voltage fluctuations at the terminals of the resistor. These fluctuations are governed by Planck's black body radiation law and have a zero mean but a nonzero rms value. At microwave frequencies the Rayleigh–Jeans approximation applies ($hf \ll kT$) such that the mean square noise voltage generated by a noisy resistor is

$$\langle v_{th}^2 \rangle = 4kTBR$$

Likewise, we may calculate the mean square noise current generated by a noisy resistor as

$$\langle i_{th}^2 \rangle = \frac{4kTB}{R}$$

The models for the noisy resistor are shown in Figure 9.19. Sometimes rms sources are shown in circuit models, and other times mean square sources are shown in the circuit models. Care must be taken to know which quantity, rms or mean square, is required by commercial software when constructing models that include these noise sources.

9.7.2 Shot Noise

Shot noise arises due to random fluctuation of electrons and photons and is manifest in the photodetection process. The mean square shot noise appearing at the photodetector output is given by the following equation where e is the electron charge, B is the detection bandwidth, and $\langle I_{PH} \rangle$ is the average photocurrent.

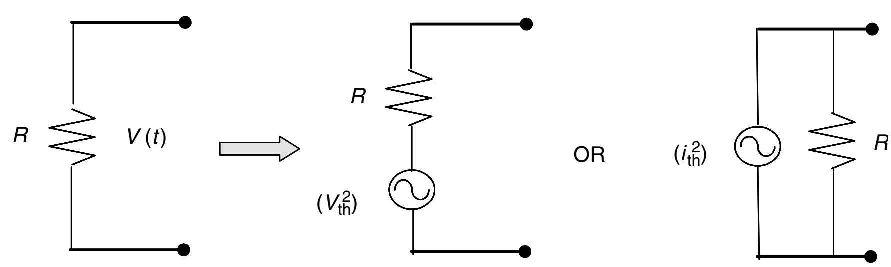


FIGURE 9.19 Noisy resistor models.

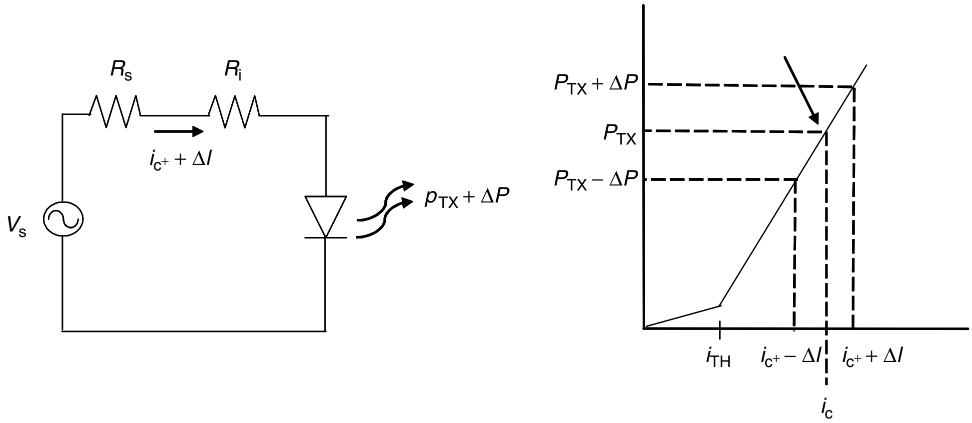


FIGURE 9.20 Figure showing the basic (RIN) process. Optical frequency dependence is not shown.

The following mean square shot noise source may be placed in parallel with the detector photocurrent source.

$$\langle i_s^2 \rangle = 2q\langle I_{PH} \rangle B$$

9.7.3 Relative Intensity Noise

Relative intensity noise, or RIN, is a function of the laser bias current and the optical frequency. Noise in the laser bias current, ΔI, results in noise in the optical power, ΔP, as shown in Figure 9.20.

RIN noise is defined is the ratio of the mean square optical power fluctuations to the average optical power. Here, the optical frequency dependence has been ignored.

$$\text{RIN} = 10 \log \left(\frac{\langle \Delta P^2 \rangle}{P_{TX}^2} \right)$$

RIN noise is translated from the optical domain to the electrical domain via the photodetection process. Therefore, the RIN is usually modeled in the electrical domain as a photocurrent noise. The following mean square noise source may be placed in parallel with the detector photocurrent source.

$$\langle i_{RIN}^2 \rangle = \frac{\langle I_{PH}^2 \rangle}{2} 10^{\text{RIN}/10} B$$

9.7.4 Noise Figure for the Directly Modulated Fiber-Optic Link

The NF for the directly modulated fiber-optic link may be determined using the model shown in Figure 9.21. It is assumed that the link is operating well below the 3 dB frequency of the photodetector so that the detector capacitance may be ignored.

The total added noise may be expressed as the summation of the added thermal noise due to the series combination of the laser resistance and series matching network, the added thermal noise due to the load resistance, and the added effects of shot noise and RIN.

$$N_{add} = N_{th-laser} + N_{th-load} + N_{sh} + N_{RIN}$$

Superposition may be applied to find the effect of each added noise term individually. Under the assumption that the laser is matched to the source via a series matching resistor (i.e., $R_M + R_i = R_s$), the laser and

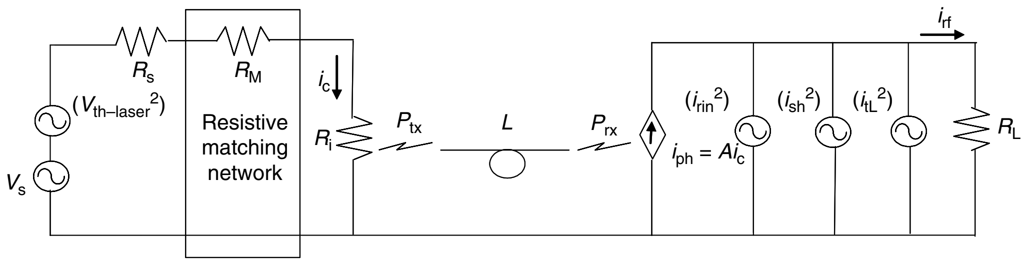


FIGURE 9.21 Circuit model for the calculation of analog fiber-optic link noise figure. This model may be used for both directly modulated and externally modulated links.

matching network will produce an input noise that is equivalent to the available noise. The added noise at the detector is this noise multiplied by the link gain.

$$N_{\text{th-laser}} = kT_0BG$$

The noise power delivered to the load due to the other sources is simply the mean square value of the noise source multiplied by the load resistor. Therefore, the total added noise is

$$N_{\text{add}} = kT_0BG + \left\{\left\langle i_{\text{tL}}^2\right\rangle + \left\langle i_{\text{sh}}^2\right\rangle + \left\langle i_{\text{RIN}}^2\right\rangle\right\}R_{\text{L}}$$

Substituting this into the definition for NF, we get

$$\text{NF} = 1 + 1 + \frac{\left\{\left\langle i_{\text{tL}}^2\right\rangle + \left\langle i_{\text{sh}}^2\right\rangle + \left\langle i_{\text{RIN}}^2\right\rangle\right\}R_{\text{L}}}{GkT_0B}$$

Substituting for the directly modulated link gain and the mean square noise values we get

$$\text{NF} = 2 + \left\{\frac{4}{R_{\text{L}}} + \frac{\langle I_{\text{ph}}^2\rangle 10^{\text{RIN}/10}}{2kT_0} + \frac{2q\langle I_{\text{ph}}^2\rangle}{kT_0}\right\}\frac{(R_{\text{M}} + R_{\text{i}})}{A^2}$$

Under the assumption that the link is resistively matched, the NF becomes

$$\text{NF} = 2 + \left\{4 + \frac{\langle I_{\text{ph}}^2\rangle 10^{\text{RIN}/10}R_{\text{L}}}{2kT_0} + \frac{2q\langle I_{\text{ph}}\rangle R_{\text{L}}}{kT_0}\right\}\frac{1}{A^2}$$

Thus, the lowest theoretical NF for a resistively matched fiber-optic link is 2 or 3 dB.

These link gain and NF models may be implemented in an RF simulator to predict link performance. Figure 9.22 shows the measured and simulated link gain and NF for optical losses from 0 to 15 dB. In this particular simulation, both pre and post link amplification was used. Good agreement between simulated and experimental values was obtained using the gain and NF models presented.

9.7.5 Fiber-Optic Link Dynamic Range

The link dynamic range is measured using a standard two-tone test. The dynamic range of the link is typically limited by the linearity of the optical transmitter, although, optical nonlinearities and photodetector linearity are also important. There are several definitions for dynamic range including CDR and SFDR. For externally modulated fiber-optic links employing the MZM, closed-form expressions may be derived for dynamic range assuming that the MZM limits the link linearity. For most other types of links,

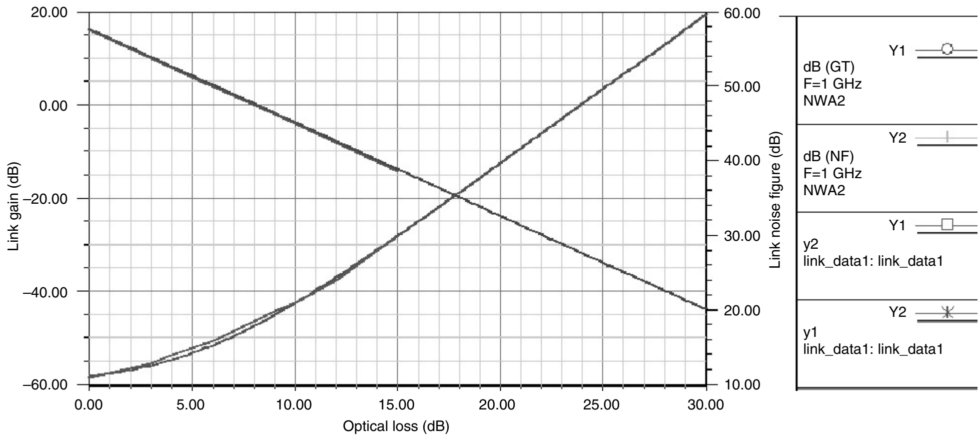


FIGURE 9.22 Measured and simulated link gain as a function of optical loss. Pre-Amp ($G = 22$ dB; NF = 2 dB); Post Amp ($G = 13.8$ dB; NF = 2 dB); $s = 0.3$ W/A; $r = 0.8$ A/W; RIN = −154 dB; $i_L = 23$ mA; optical coupling loss = 0.6 dB; $I_{ph} = 4.8$ mA.

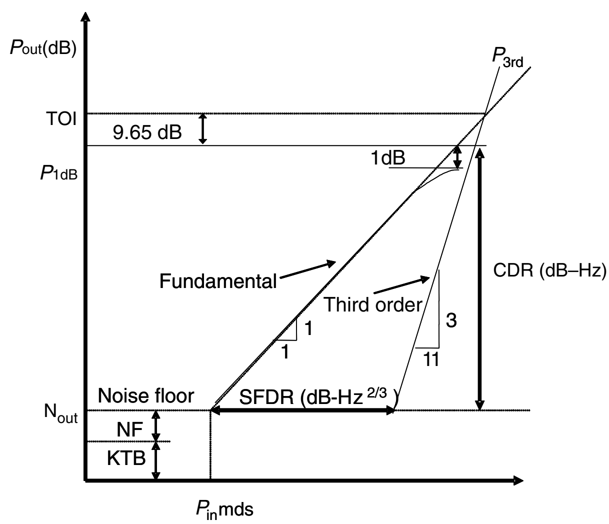


FIGURE 9.23 Plot showing analog fiber-optic link dynamic range definitions assuming third-order nonlinearities are dominant.

including the directly modulated link and the externally modulated link using the EAM, a power series expansion of the transfer function nonlinearity can be used to predict the link dynamic range in a fashion analogous to that used for power amplifiers. Figure 9.23 shows a plot of the input and output signals of a fiber link for a two-tone test assuming that the dominant nonlinearity is third order.

9.7.5.1 Compression Dynamic Range (dB-Hz)

The CDR is specified as the range of input signal for which the output is above the noise floor and compressed by less than 1 dB. Bandwidth normalization is $10\log(B)$.

$$\mathrm{CDR} = \frac{1.259 P_{\mathrm{out}_{1\mathrm{dB}}}}{N_{\mathrm{out}}} = \frac{G P_{\mathrm{in}_{1\mathrm{dB}}}}{kTBG \times \mathrm{NF}} = \frac{P_{\mathrm{in}_{1\mathrm{dB}}}}{kTB \times \mathrm{NF}}$$

$$\mathrm{CDR(dB\text{-}Hz)} = P_{\mathrm{in}_{1\mathrm{dB}}}(\mathrm{dB}) - \mathrm{NF(dB)} - 174 - 10\log(B)$$

9.7.5.2 Spurious-Free Dynamic Range (dB-Hz$^{2/3}$)

The SFDR is the range of input signal for which the fundamental output is above the noise floor and the third-order intermodulation products are below the noise floor. Bandwidth normalization is $2/3\log(B)$.

$$\text{SFDR} = \left(\frac{\text{TOI}}{N_{\text{out}}}\right)^{2/3} = \left(\frac{\text{TOI}}{kTBG \times \text{NF}}\right)^{2/3}$$

$$\text{SFDR}(\text{dB-Hz}^{2/3}) = \frac{2}{3}[\text{TOI} - G - N_{\text{in}}]$$

SFDR can also be defined for a linearized system. In the case where the linearization eliminates the third-order nonlinearity, the fifth-order nonlinearity is of concern for sub-octave systems (dB-Hz$^{4/5}$). Bandwidth normalization is $4/5\log(B)$.

Closed-form expressions can be obtained for the CDR and SFDR of externally modulated fiber-optic links using the MZM [57]. The CDR is analyzed by substituting a single frequency signal into the MZM transfer function and performing a Bessel series expansion. The amplitude of the AC input signal that results in 1 dB compression at the output is

$$v_{s_{1\text{dB}}} = \frac{1.0095\,V_\pi}{\pi}$$

The input power corresponding to the 1 dB compression point is

$$P_{\text{in}_{1\text{dB}}} = \left(\frac{1.0095\,V_\pi}{\pi}\right)^2 \frac{1}{2R_{\text{M}}}$$

The output power at the 1 dB compression point is given by

$$P_{\text{out}_{1\text{dB}}} = \frac{P_{\text{in}_{1\text{dB}}} G_{\text{mzm}}}{1.2589} = \left(\frac{1.0095\,V_\pi}{\pi}\right)^2 \frac{g_{\text{mzm}}^2 R_{\text{L}}}{2\,(1.2589)}$$

resulting in a CDR of

$$\text{CDR} = \frac{1.259 P_{\text{out}_{1\text{dB}}}}{N_{\text{out}}}$$

The spurious-free dynamic range analysis for the externally modulated link using the MZM can be performed by inserting a two-tone excitation into the MZM transfer function.

$$P_0 = L_{\text{M}}\frac{P_{\text{i}}}{2}\left\{1 + \sin\left(\frac{\pi\,V_\pi + \pi\,v_{\text{m}}\sin(\omega_1 t) + \pi\,v_{\text{m}}\sin(\omega_2 t)}{V_\pi}\right)\right\}$$

A Bessel serious expansion may be performed to determine the amplitudes of the fundamental (ω_1, ω_2) and third-order $(2\omega_2\text{-}\omega_1, 2\omega_1\text{-}\omega_2)$ frequency components. The SFDR is given in terms of a modulation index, m_{int}, that corresponds to equal power levels of fundamental and third-order intermodulation distortion products (e.g., at the third-order intercept point). This index is an extrapolated modulation index (similar to the third-order intercept itself) that would result in the fundamental and third-order frequency components having equal power.

$$m_{\text{int}} = \frac{\sqrt{32}}{\pi\left(1 - \frac{2|V_{\text{B}}|}{V_\pi}\right)}$$

The input power corresponding to the third-order intercept is

$$P_{\text{in}_{\text{TOI}}} = \left[\frac{m_{\text{int}}}{2}\left(V_\pi - 2|V_{\text{B}}|\right)\right]^2 \frac{1}{2R_{\text{M}}}$$

The output third-order intercept is given by

$$P_{\text{out}_{\text{TOI}}} = P_{\text{in}_{\text{TOI}}} G_{\text{mzm}} = \left[\frac{m_{\text{int}}}{2}\left(V_\pi - 2|V_{\text{B}}|\right)\right]^2 \frac{g_{\text{m}}^2 R_{\text{L}}}{2}$$

Finally, the SFDR can be found as

$$\text{SFDR} = \left(\frac{P_{\text{out}_{\text{TOI}}}}{N_{\text{out}}}\right)^{2/3}$$

Figure 9.24 shows the 1 dB compression characteristics for an externally modulated fiber-optic link with the following parameters.

Figure 9.25 shows the simulated two-tone test for the same link.

Table 9.2 shows the theoretical and simulated performance of the link compression point and third-order intercept.

9.8 Analog Fiber-Optic Link Analysis Summary

This chapter presented an analysis of link gain, NF, and dynamic range for the resistively matched topology of directly modulated and externally modulated fiber-optic links. Reactive impedance matching may be used to improve gain and/or NF at the expense of link bandwidth [16]. Currently, the performance of fiber-optic links [58] is adequate for many analog applications including CATV, radio over fiber, antenna remoting, and various test and measurement applications. Various link linearization approaches can improve the performance even further and optical amplification can be used to reduce the effects of optical power loss [59]. However, there are still challenges associated with achieving the performance required

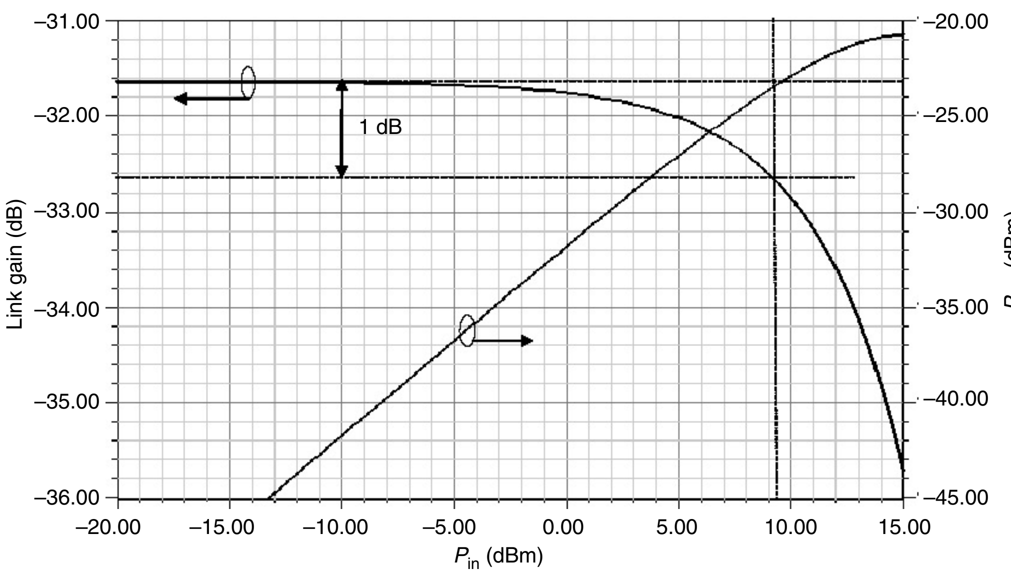


FIGURE 9.24 Simulated 1 dB compression characteristics for an externally modulated link using an MZM.

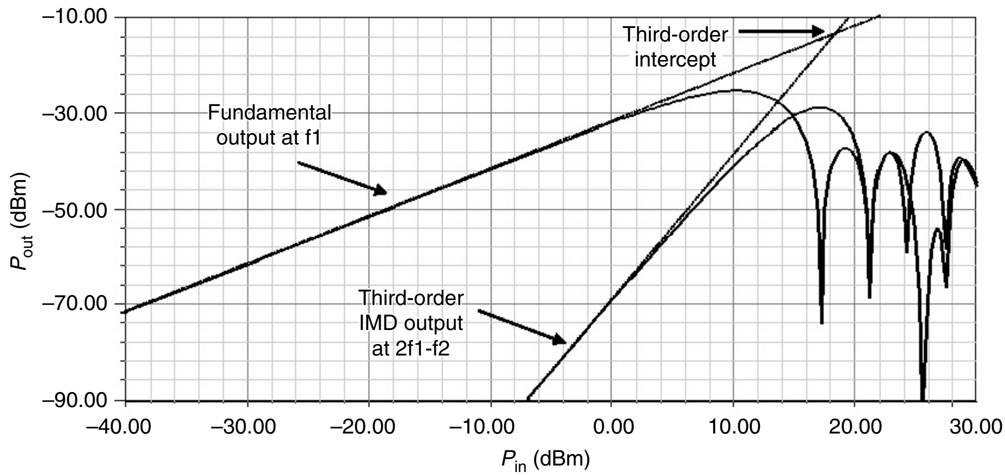


FIGURE 9.25 Simulated third-order intercept for an externally modulated link using an MZM.

TABLE 9.2 Theoretical and Simulated Externally Modulated Link Parameters (Simulations Use the Models Presented in this Chapter)

	Theoretical	Simulated
G_{mzm}(dB)	−31.57	−31.63
P_{in1dB}(dBm)	9.68	9.25
P_{out1dB}(dBm)	−22.90	−23.40
P_{inTOI}(dBm)	18.6	18.5
P_{outTOI}(dBm)	−13.0	−13.5

by demanding high dynamic range applications such as radar signal distribution and the development of new link topologies remains an active research area. Class AB [60], class E [61], traveling-wave [62], and balanced EAM links [63] have all been developed recently. There is also a high level of interest in the use coherent fiber-optic links [64,65] to achieve high dynamic range requirements.

References

1. Paolella, A., LaFond, K., Borlando, J., Aviles, M., Simpkins, C., "Dense wavelength division multiplexing photonic transport for wireless access systems," *Optical Fiber Communication Conference and Exhibit, 2002*. OFC 2002, Anaheim, CA, pp. 768–769, March 17–22, 2002.
2. Paolella, A.C., Joshi, A.M., Bauerle, A., "Photonic modules for millimeter wave communication systems," *Microwave Photonics, 2000*. MWP 2000. International Topical Meeting, St. Louis, Missouri, pp. 233–236, 2000.
3. Paolella, A., *A New High Efficiency, High Power Transmitter System for Wireless Communications*, Milcom 2005, Atlantic City, NJ, October 20, 2005.
4. Paolella, A., et al., "Fiberoptic dual delay line for a multimode radar test target simulator," *Microwave Symposium Digest, IEEE MTT-S International*, Vol. 2, Atlanta, GA, pp. 1059–1062, June 14–18, 1993.
5. Paolella, A., *The Sidelighter: A New Dead-Zone Free Distance to Fault Fiber Optic Cable Measurement System, 2006 IEEE LEOS AVFOP*, Annapolis, MD, September 12–14, 2006.
6. Seeds, A.J., "Microwave photonics," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 50, No. 3, pp. 877–887, March 2002.

7. Yao, X.S., Maleki, L., Eliyahu, D., "Progress in the opto-electronic oscillator—a ten year anniversary review," *Microwave Symposium Digest, 2004 IEEE MTT-S International*, Vol. 1, pp. 287–290, June 6–11, 2004.
8. Minasian, R.A., "Photonic signal processing of microwave signals," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 54, No. 2, pp. 832–846, February 2006.
9. Lindsay, A.C., Knight, G.A., Winnall, S.T., "Photonic mixers for wide bandwidth RF receiver applications," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 43, No. 9, pp. 2311–2317, September 1995.
10. Soref, R., "Voltage-controlled optical/RF phase shifter," *Journal of Lightwave Technology*, Vol. 3, No.5, pp. 992–998, October 1985.
11. Frigyes, I., Seeds, A.J., "Optically generated true-time delay in phased-array antennas," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 43, No. 9, pp. 2378–2386, September 1995.
12. Tonda-Goldstein, S., Dolfi, D., Monsterleet, A., Formont, S., Chazelas, J., Jean-Pierre Huignard, "Optical signal processing in Radar systems," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 54, No. 2, pp. 847–853, February 2006.
13. Way, W.I., *Broadband Hybrid Fiber/Coax Access System Technologies*, Academic Press, San Diego, CA, 1999.
14. Cox, C.H., III, Ackerman, E.I., Betts, G.E., Prince, J.L., "Limits on the performance of RF-over-fiber links and their impact on device design," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 54, No. 2, pp. 906–920, February 2006.
15. Jalali, B., Paniccia, M., Reed, G., "Silicon photonics," *IEEE Microwave Magazine*, Vol. 7, No. 3, pp. 58–68, June 2006.
16. Cox, C.H., *Analog Optical Links—Theory and Practice*, Cambridge University Press, Cambridge, UK, 2004.
17. Chang, W.S.C., *RF Photonic Technology in Optical Fiber Links*, Cambridge University Press, Cambridge, UK, 2002.
18. Welch, D.F., "A brief history of high-power semiconductor lasers," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 6, No. 6, pp. 1470–1477, November/December 2000.
19. Suematsu, Y., Arai, S., "Single-mode semiconductor lasers for long-wavelength optical fiber communications and dynamics of semiconductor lasers," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 6, No. 6, pp. 1436–1449, November/December 2000.
20. Tucker, R.S., Pope, D.J., "Microwave circuit models of semiconductor injection lasers," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 83, No. 3, pp. 289–294, March 1983.
21. Hafskjaer, L., Sudbo, A., "Modeling of the frequency modulation response of semiconductor diode lasers," *IEEE Journal of Quantum Electronics*, Vol. 24, No. 4, pp. 625–634, April 1988.
22. Majewski, M.L., Novak, D., "Method for characterization of intrinsic and extrinsic components of semiconductor laser diode circuit model," *IEEE Microwave and Guided Wave Letters [see also IEEE Microwave and Wireless Components Letters]*, Vol. 1, No. 9, pp. 246–248, September 1991.
23. Sum, K.C., Gomes, N.J., "Microwave-optoelectronic modeling approaches for semiconductor lasers," *IEE Proceedings in Optoelectronics*, Vol. 145, No. 3, pp. 141–146, June 1998.
24. Hafskjaer, L., Sudbo, A., "Modeling of the frequency modulation response of semiconductor diode lasers," *IEEE Journal of Quantum Electronics*, Vol. 24, No. 4, pp. 625–634, April 1988.
25. Iezekiel, S., Snowden, C.M., Howes, M.J., "Nonlinear circuit analysis of harmonic and intermodulation distortions in laser diodes under microwave direct modulation," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 38, No. 12, pp. 1906–1915, December 1990.
26. Jianjun Gao, Xiuping Li, Flucke, J., Boeck, G., "Direct parameter-extraction method for laser diode rate-equation model," *Journal of Lightwave Technology*, Vol. 22, No. 6, pp. 1604–1609, June 2004.
27. Kato, K., "Ultrawide-band/high-frequency photodetectors," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 7, pp. 1265–1281, July 1999.
28. Williams, K.J., Esman, R.D., Dagenais, M., "Nonlinearities in p-i-n microwave photodetectors," *Journal of Lightwave Technology*, Vol. 14, No. 1, pp. 84–96, January 1996.
29. Williams, K.J., Tulchinsky, D.A., Boos, J.B., Doewon Park, Goetz, P.G., "High-power photodiodes," *2006 Digest of the LEOS Summer Topical Meetings*, Quebec City, Canada, pp. 50–51, July 17–19, 2006.
30. Kubota, K., Noda, J., Mikami, O., "Traveling wave optical modulator using a directional coupler LiNbO3waveguide," *IEEE Journal of Quantum Electronics*, Vol. 16, No. 7, pp. 754–760, July 1980.

31. Wooten, E.L., Kissa, K.M., Yi-Yan, A., Murphy, E.J., Lafaw, D.A., Hallemeier, P.F., Maack, D., Attanasio, D.V., Fritz, D.J., McBrien, G.J., Bossi, D.E., "A review of lithium niobate modulators for fiber-optic communications systems," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 6, No. 1, pp. 69–82, January/February 2000.
32. Chen, D., Fetterman, H.R., Chen, A., Steier, W.H., Dalton, L.R., Wang, W., Shi, Y., "Demonstration of 110 GHz electrooptic polymer modulators," *Applied Physics Letters*, Vol. 70, pp. 3335–3337, 1997.
33. Erman, M., Jarry, P., Gamonal, R., Autier, P., Chane, J.-P., Frijlink, P., "Mach-Zehnder modulators and optical switches on III-V semiconductors," *Journal of Lightwave Technology*, Vol. 6, No. 6, pp. 837–846, June 1988.
34. Alferness, R., Korotky, S., Marcatili, E., "Velocity-matching techniques for integrated optic traveling wave switch/modulators," *IEEE Journal of Quantum Electronics*, Vol. 20, No. 3, pp. 301–309, March 1984.
35. Farwell, M.L., Chang, W.S.C., Huber, D.R., "Increased linear dynamic range by low biasing the Mach-Zehnder modulator," *IEEE Photonics Technology Letters*, Vol. 5, No. 7, pp. 779–782, July 1993.
36. Jemison, W.D., Kreuzberger, A.J., Funk, E., "Microwave photonic vector modulator for high-speed wireless digital communications," *IEEE Microwave and Wireless Components Letters [see also IEEE Microwave and Guided Wave Letters]*, Vol. 12, No. 4, pp. 125–127, April 2002.
37. Katz, A., Jemison, W., Kubak, M., Dragone, J., "Improved radio over fiber performance using predistortion linearization," *2003 IEEE MTT-S International Microwave Symposium Digest*, Vol. 2, pp. 1403–1406, June 8–13, 2003.
38. Meng, X.J., Yacoubian, A., Bechtel, J.H., "Electro-optical predistortion technique for linearisation of Mach-Zehnder modulators," *Electronics Letters*, Vol. 37, No. 25, pp. 1545–1547, December 6, 2001.
39. Skeie, H., Johnson, R., "Linearization of electrooptic modulators by a cascade coupling of phase modulating electrodes," *Proceedings of SPIE—The International Society for Optical Engineering*, Vol. 1583, pp. 153–164, March 1991.
40. Korotky, S., DeRidder, R., "Dual parallel modulation schemes for low-distortion analog optical transmission," *IEEE Journal on Selected Areas in Communication*, Vol. 8, pp. 1377–1381, September 1990.
41. Farwell, M.L., Lin, Z.-Q., Wooten, E., Chang, W.S.C., "An electrooptic intensity modulator with improved linearity," *IEEE Photonics Technology Letters*, Vol. 3, No. 9, pp. 792–795, September 1991.
42. Johnson, L.M., Roussell, H.V., "Linearization of an interferometric modulator at microwave frequencies by polarization mixing," *IEEE Photonics Technology Letters*, Vol. 2, No. 11, pp. 810–811, November 1990.
43. Ackerman, E.I., "Broad-band linearization of a Mach-Zehnder electrooptic modulator," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 12, pp. 2271–2279, December 1999.
44. Li, G.L., Sun, C.K., Pappert, S.A., Chen, W.X., Yu, P.K.L., "Ultrahigh-speed traveling-wave electroabsorption modulator-design and analysis," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, No. 7, pp. 1177–1183, July 1999.
45. Johansson, L.A., Akulova, Y.A., Fish, G.A., Coldren, L.A., "High optical power electroabsorption waveguide modulator," *Electronics Letters*, Vol. 39, No. 4, pp. 364–365, February 20, 2003.
46. Li, G.L., Yu, P.K.L., Chang, W.S.C., Loi, K.K., Sun, C.K., Pappert, S.A., "Concise RF equivalent circuit model for electroabsorption modulators," *Electronics Letters*, Vol. 36, No. 9, pp. 818–820, April 27, 2000.
47. Betts, G.E., Xie, X.B., Shubin, I., Chang, W.S.C., Yu, P.K.L., "Gain limit in analog links using electroabsorption modulators," *IEEE Photonics Technology Letters*, Vol. 18, No. 19, pp. 2065–2067, October 1, 2006.
48. Loi, K.K., Hodiak, J.H., Mei, X.B., Tu, C.W., Chang, W.S.C., "Linearization of 1.3-μm MQW electroabsorption modulators using an all-optical frequency-insensitive technique," *IEEE Photonics Technology Letters*, Vol. 10, No. 7, pp. 964–966, July 1998.
49. Yu-Dong Bae, Byung-Kwon Kang, Byeonghoon Park, Sang-Moon Lee, Young Hyun Kim, Hong-Kwon Kim, Mun-Kue Park, In Kim, Jang, D.-H., "Operation of 1550-nm electroabsorption Modulated laser at 40/spl deg/C for 10-gb/s, 40-km transmission," *IEEE Journal of Selected Topics in Quantum Electronics*, Vol. 11, No. 1, pp. 135–140, January–February 2005.

50. Hyun-Do Jung, Sang-Yun Lee, Sang-Kook Han, "Gain saturated SOA cascaded EML analog optical transmitter," *2002 International Topical Meeting on Microwave Photonics*, Awaji, Japan, pp. 205–208, November 5–8, 2002.
51. Kurniawan, T., Nirmalathas, A., Lim, C., Novak, D., Waterhouse, R., "Performance analysis of optimized millimeter-wave fiber radio links," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 54, No. 2, pp. 921–928, February 2006.
52. Schmuck, H., "Comparison of optical millimetre-wave system concepts with regard to chromatic dispersion," *Electronics Letters*, Vol. 31, No. 21, pp. 1848–1849, October 12, 1995.
53. Smith, G.H., Novak, D., Ahmed, Z., "Technique for optical SSB generation to overcome dispersion penalties in fibre-radio systems," *Electronics Letters*, Vol. 33, No. 1, pp. 74–75, January 2, 1997.
54. Smith, G.H., Novak, D., "Broad-band millimeter-wave (38 GHz) fiber-wireless transmission system using electrical and optical SSB modulation to overcome dispersion effects," *IEEE Photonics Technology Letters*, Vol. 10, No. 1, pp. 141–143, January 1998.
55. Funk, E.E., Urick, V.J., Strutz, S.J., Dexter, J.L., Williams, K.J., "110 km 256-QAM digital microwave over fiber link," *2003 IEEE MTT-S International Microwave Symposium Digest*, Vol. 1, pp. 269–272, June 8–13, 2003.
56. Schmuck, H., "Effect of polarisation-mode-dispersion in fibre-optic millimetre-wave systems," *Electronics Letters*, Vol. 30, No. 18, pp. 1503–1504, September 1, 1994.
57. Daryoush, A.S., Ackerman, E., Samant, N.R., Wanuga, S., Kasemset, D., "Interfaces for high speed fiber-optic links: analysis and experiment," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 39, No. 12, pp. 2031–2044, December 1991.
58. Ackerman, E.I., Cox, C.H., "RF fiber-optic link performance," *IEEE Microwave Magazine*, Vol. 2, No. 4, pp. 50–58, December 2001.
59. Sabido, D.J.M., Kazovsky, L.G., "Dynamic range of optically amplified RF optical links," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 49, No. 10, pp. 1950–1955, October 2001.
60. Bull, J.D., Darcie, T.E., Jinye Zhang, Kato, H., Jaeger, N.A.F., "Broadband class-AB microwave-photonic link using polarization modulation," *IEEE Photonics Technology Letters*, Vol. 18, No. 9, San Francisco, pp. 1073–1075, May 1, 2006.
61. Jemison, W.D., Paolella, A.C., "A novel class E analog fiber optic link with RF power gain and high DC-RF conversion efficiency, *2006 IEEE MTT-S International Microwave Symposium*, pp. 14–17, June 2006.
62. Hashim, H.H., Iezekiel, S., "Traveling-wave microwave fiber-optic links," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 54, No. 2, pp. 951–958, February 2006.
63. Mathai, S., Cappelluti, F., Jung, T., Novak, D., Waterhouse, R.B., Sivco, D., Cho, A.Y., Ghione, G., Wu, M.C., "Experimental demonstration of a balanced electroabsorption modulated microwave photonic link," *IEEE Transactions on Microwave Theory and Techniques*, Vol. 49, No. 10, pp. 1956–1961, October 2001.
64. Kalman, R.F., Fan, J.C., Kazovsky, L.G., "Dynamic range of coherent analog fiber-optic links," *Journal of Lightwave Technology*, Vol. 12, No. 7, pp. 1263–1277, July 1994.
65. Li, Y., Jemison, W.D., Herczfeld, P.R., Rosen, A., "Coherent phase modulated (PM) fiber optic link design," *2006 IEEE MTT-S International Microwave Symposium*, San Francisco, pp. 1943–1946.

10

Satellite Communications Systems

Ramesh K. Gupta
Mobile Satellite Ventures

10.1 Introduction

The launch of commercial communications satellite services in the early 1960s ushered in a new era in international telecommunications that has affected every facet of human endeavor. Although communications satellite systems were first conceived to provide reliable telephone and facsimile services among nations around the globe, today satellites provide worldwide TV channels (24 h a day), global messaging services, positioning information, communications from ships and aircraft, communications to remote areas, disaster relief and emergency services on land and sea, personal communications, and high-speed data services including Internet access. The percentage of voice traffic being carried over satellites—which stood at approximately 70% in the 1980s—has been rapidly declining [1] as new video, data and IP services are added over existing satellite networks.

The demand for fixed satellite services (FSS) and capacity continues to grow. Rapid deployment of private networks using very small aperture terminals (VSATs)—which interconnect widely dispersed corporate offices, manufacturing, supply, and distribution centers—has contributed to this growth. The use of VSATs, which employ a low-cost terminal with a relatively small aperture antenna (1–2 m), has steadily grown around the world. These terminals are inexpensive and easy to install, and are linked together with a large central hub station, through which all the communications take place. The majority of these VSAT applications previously used data rates of 64 or 128 kb/s. However, the demand for higher rate VSAT services (with data rates of 2 Mb/s and higher for corporate users) has been growing due to emerging multimedia services [2], largely fueled by the growth in the data and Internet traffic.

Internet and video distribution services are a major and rapidly increasing component of satellite services. Internet and IP-related services (largest growth segments) being offered by many satellite service providers include IP videoconferencing, broadcasting, video steaming, IP Multicasting [3,4], IPTV, Cellular backhaul, and also Voice over IP (VoIP). Growth in international trade, trends in manufacturing and service outsourcing, growth in data traffic, and significant drops in prices of desk tops, portable computers, and memory devices are all expected to further contribute to the trend. Key attributes of satellite networks, namely—global coverage, provision of "instantaneous infrastructure" and services, broadcasting and multicasting capabilities of video and data traffic are well recognized as cost-effective solutions for provision of wide area services. In developing countries satellites are used to bridge the "digital divide" among communities and population centers. In the developed world, satellites provide services to the edge of the telecommunications networks and fill the voids in areas where terrestrial fiber and wireless infrastructure are not economically feasible and/or are temporarily disabled because of natural or man-made disasters. The use of communications satellites for disaster recovery was amply demonstrated during major disasters such as 2004 tsunami in Asia [5] and also during the recovery operations (2005) in the aftermath of hurricanes Katrina and Rita in the United States. Because of these disasters, satellite-based emergency services have won praise from regulatory agencies and have also become one of the key drivers for emerging convergence between satellite and wireless services.

In the past two decades, wireless communications has emerged as one of the fastest growing services. Cellular wireless services, which have been leading the growth over the past decade, achieved cumulative annual growth rates in excess of 30% [6]. As the growth rates for cellular phones in Europe and United States somewhat leveled off, Asia became a very attractive growth market with India and China leading this growth with the number of cellular phones already taking over the number of fixed lines. To complement this rapidly emerging wireless telecommunications infrastructure around the world, a number of satellite-based global personal communications systems (PCS) [7–9] were conceived and deployed. Examples include low earth orbit (LEO) systems such as Iridium and Globalstar, medium earth orbit (MEO) systems such as ICO Global, and geostationary (GEO) systems such as Asia Cellular Satellite System (ACeS) and Thuraya. These systems were largely designed to provide narrowband voice and data services to small handheld terminals. Many of these systems experienced problems with market penetration because of their significant system costs and also rapid deployment of cellular infrastructure in their target markets in which they were trying to compete. These satellite systems are discussed in significant detail in the next chapter and, therefore, will not be discussed further.

In parallel with these developments, rapid growth in Internet and multimedia traffic has created an exponential increase in the demand for transmission bandwidth. For example, Internet and other data services have been contributing more than 70% of new revenues for major satellite operators [3]. The share of data and IP traffic via satellites is expected to grow as new service applications emerge and as the number of Internet users continues to grow. As experience in Korea and Japan has shown, the largest growth contributors have been the mobile Internet subscribers, expected to exceed fixed Internet subscribers by the year 2010. According to International Telecommunications Union (ITU) estimates, provision of high-speed data services and the multimedia traffic are expected to contribute more than five times the growth in traffic in 2010 as compared to traffic recorded in 1999 [10]. The new services for home or business users include high-speed Internet access, e-commerce, home shopping, telecommuting, telemedicine, distance education, teleconferencing, and other voice, data, and image-based communications services. At present, most satellite-based broadband services are being offered at Ku-band, using broadband VSAT-type terminals. Additional systems were proposed for the Ka- and V-bands [11,12]. Since 1995, 14 Ka-band (30/20-GHz) and 16 V-band (50/40-GHz) satellite systems filings were made to the Federal Communications Commission (FCC) in the United States in response to the Ka- and V-band frequency allotments. However, very few of these proposed systems could come to full fruition because of limited availability of investments as there has been a continuing shift in telecommunications markets, services, and investment environment, particularly the United States. This chapter reviews the evolution of communications satellite systems, addresses various service offerings, and describes the characteristics of deployed Ka-band broadband satellite systems and also some of the emerging hybrid satellite systems.

10.2 Evolution of Communications Satellites

Communications satellites have experienced an explosive growth in traffic over the past several decades due to a rapid increase in global demand for voice, video, and data traffic and more recently for multimedia and IP services. For example, INTELSAT I (Early Bird) satellite, launched in 1965 by the International Telecommunications Satellite Organization's [5], carried only one wideband transponder operating at C-band frequencies and could support 240 voice circuits. A large 30-m diameter Earth Station antenna was required to close the link. In contrast, the communications payload of the INTELSAT VI series of satellites, launched from 1989 onward, provided 50 distinct transponders operating over both C- and Ku-bands [13], providing the traffic-carrying capacity of approximately 33,000 telephone circuits. By using digital compression and multiplexing techniques, INTELSAT VI capacity was enhanced to support 120,000 two-way telephone channels and three television channels [14]. The effective isotropic radiated power (EIRP) of these and follow-on satellites is sufficient to allow use of much smaller earth terminals at Ku-band (E1: 3.5 m) and C-bands (F1: 5 m), in addition to other larger earth stations. Several technological innovations including the light weight, highly reliable GaAs MMIC components/subsystems and the use of narrow spot beams with higher EIRP have contributed to this increase in the number of active transponders required to satisfy growing traffic demand. A range of voice, data, and IP services are available today through communications satellites to fixed as well as mobile users. In general, the satellite services may be classified into three broad categories—FSS, Direct Broadcast Services, and Mobile Satellite Services—which are discussed in further detail in the following sections.

10.2.1 Fixed Satellite Services

With fixed satellite services [FSS], signals are relayed between the satellite and relatively large fixed earth stations. Terrestrial landlines and fiber are used to connect voice, television, data and IP traffic to these earth stations. Figure 10.1 shows the FSS configuration. The FSS were the first services to be developed for global communications. COMSAT was formed in 1962 through an act of U.S. Congress with a charter to develop satellite services in the United States and around the globe. Subsequently, INTELSAT with headquarters in Washington, DC, was formed to provide these services with COMSAT serving as the U.S. signatory. With satellites operating over the Atlantic, Indian, and Pacific Ocean regions, INTELSAT initially provided global communications service through its signatories (largely Government PTTs). With privatization of the communications industry in the United States and many countries around the world, many new regional and global satellite operators have emerged, capable of providing end-to end satellite services. For example, with the recent merger of a privatized INTELSAT and PanAmSat, INTELSAT has emerged as the largest operator of FSS with more than 50 satellites in orbit and providing end-to-end telecommunications services around the globe.

Several regional satellite systems have also developed around the world for communications among nations within a continent or with common regional interests. As an example, these systems include Arabsat, JSAT, and Asiasat among others. For large and geographically diverse countries like the United States, Canada, India, China, Australia, and Indonesia, the FSS are being used to establish communications links for domestic and also regional customers. The growth in communications traffic and globalization of world economies has resulted in privatization of telecommunications services and also acquisition of satellite companies, leading to further consolidation of satellite operators and industry. According to International Telecommunication Union (ITU), estimated 85% of all communications traffic was being provided under competitive market conditions in 2005, as compared to just 35% in 1990 [6]. This trend

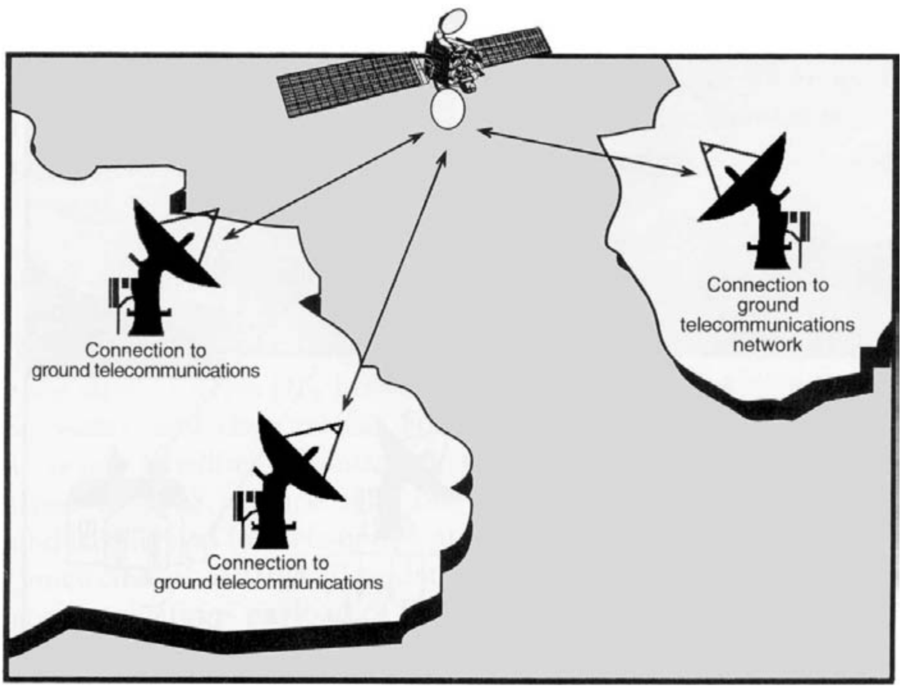


FIGURE 10.1 Fixed satellite communications services (FSS).

is expected to continue resulting in a variety of services for the users at prices driven by economic and market forces. As discussed later, the FSS satellite services are rapidly emerging toward a hybrid network model with satellite operators providing integrated end-to-end value-added services rather than leased "bandwidth only" services.

10.2.2 Direct Broadcast Satellite Services

Direct broadcast satellites (DBS) use relatively high-power satellites to distribute television programs directly to subscriber homes (Figure 10.2), or to community antennas from which the signal is distributed to individual homes by cable. The use of medium-power satellites such as Asiasat has stimulated growth in the number of TV programs available in countries such as India and China. Launch of DBS services such as DIRECTV (Hughes Networks) and Dish Network (Echostar) provides multiple pay channels via an 18-in. antenna. The growth of this market has been driven by the availability of low-cost receive equipment consisting of an antenna and a low-noise block down-converter (LNB). This market segment has seen an impressive growth, with estimated over 25 million direct-to-home TV subscribers in the United States in 2005. On a worldwide basis, installed base of digital set top boxes for television video services was estimated at 65% (as compared to cable and other terrestrial delivery) in 2002 [15]. However, the growth rates of cable and terrestrial technologies have been significantly higher during the same time, partially because of lack of a suitable return path for "triple play" (video, voice, and data). Therefore, future availability of return connection through wireless or satellite is critical for sustainable growth of the DTH services.

Another area of major growth has been nationwide broadcasting of Digital Audio Radio Services (DARS). In the United States, Sirius Satellite Radio and XM Radio provide satellite-based radio services to mobile (in car) and fixed users at S-band (2.3 GHz), whereas WorldSpace provides satellite-based radio services worldwide (except United States where they are investors in XM Radio) at L-band. Majority of Sirius and XM Satellite Radio customers are automobile users who receive 70–100 music, talk show, and news channels with minimal commercials for a modest monthly subscription fee. WorldSpace provides

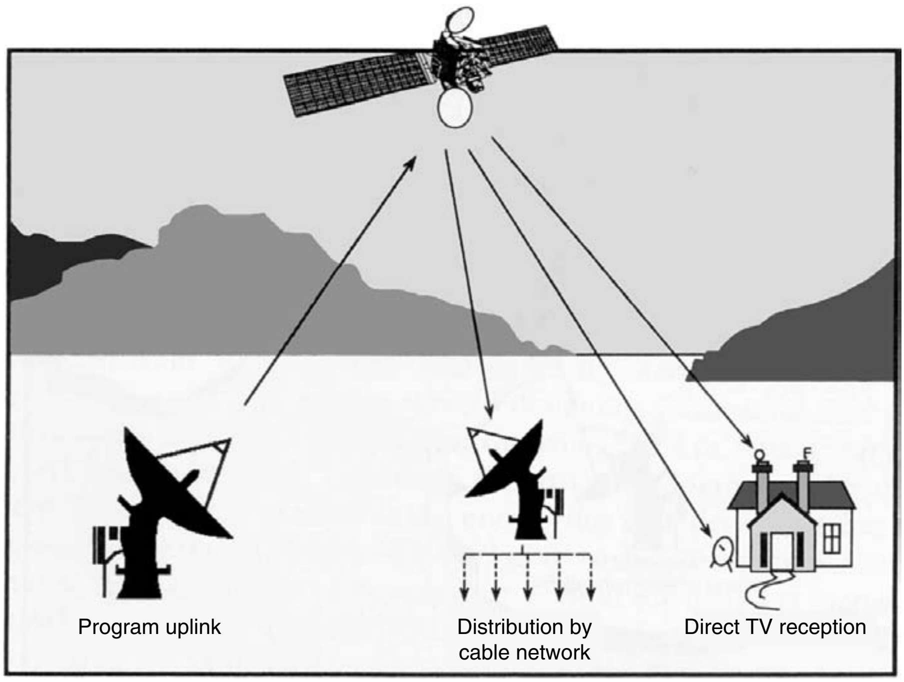


FIGURE 10.2 Direct broadcast satellite communications services (DBS). Direct-to-home (DTH) services are also offered with similar high-power Ku-band satellite systems.

more than 40 channels of radio programming to largely fixed users in Asia and Africa using their three geosynchronous satellites. XM Radio uses two geosynchronous satellites with direct line of sight coverage to the users whereas Sirius Satellite Radio uses three satellites in elliptical orbits, out of which two satellites are visible to a user. Both these systems use a large number of terrestrial repeaters to provide additional coverage in tunnels, under the bridges and in dense urban areas where buildings and high rises create satellite signal blockage and multipath effects. Orthogonal frequency division multiplexing (OFDM) frequency modulation techniques are used in these systems to minimize the impact of multipath fading.

10.2.3 Mobile Satellite Services

With mobile satellite services (MSS), communication takes place between a large fixed earth station and a number of smaller earth terminals fitted on vehicles, ships, boats, or aircraft (Figure 10.3). Originally, the International Maritime Satellite Organization (Inmarsat), headquartered in London [16], was created in 1976 to provide space segment to improve maritime communications, improve distress communications, and enhance the safety of life at sea. Subsequently, Inmarsat's mandate was extended to include aeronautical services and land-mobile services including to trucks and trains. Launch of the second generation (Inmarsat 2) satellites in early 1990s, provided L-band capacity for a minimum of 150 simultaneous maritime telephone calls using Inmarsat -A, -B, and -M terminals. With the launch of third-generation (Inmarsat 3) satellites in 1996–1998, Inmarsat mobile satellite communications capacity increased more than 10-folds by frequency reuse through seven spot beams. With an EIRP of 49 dBW, these satellites introduced the lower cost Mini-M terminals for 4.8 kbps voice, data, and fax services. Inmarsat satellite system has evolved to provision of mobile data services at 144–492 kbps using fourth-generation (inmarsat-4) satellites and broadband global area network (BGAN) terminals [17]. These satellites, first launched in 2005, increase the satellite capacity through use of a global, 19 spot and approximately 200 narrow spot beams. 64 kbps IDSN, voice, standard, and streaming IP services are available with BGAN terminals manufactured by different terminal vendors.

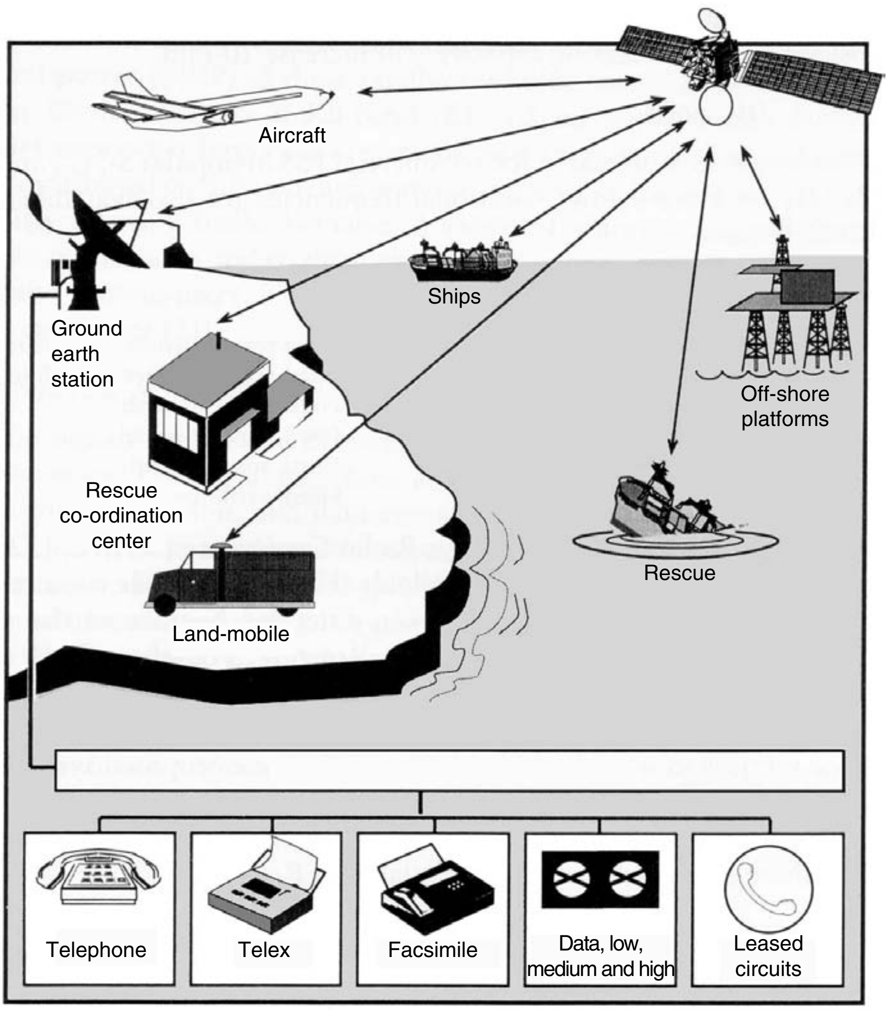


FIGURE 10.3 Mobile satellite communications services (MSS).

In the United States, the FCC granted a license to American Mobile Satellite Corporation (AMSC) [18] in 1989 to build and operate MSS satellites with L-band coverage in North America (United States, Canada, and Mexico). In 1990, AMSC and Telesat Mobile Inc (TMI) of Canada came together for joint development for MSAT-1 and MSAT-2 satellites. These satellites (1995 launch) use four of their six beams to cover mainland United States and Canada and coastal waters (up to 200 nautical miles). A fifth beam provides coverage for Alaska and Hawaii, while a sixth beam serves Mexico, Central America, and the Caribbean. L-band frequency is reused in the East and West beams, which are spatially isolated from each other. Two feeder-link earth stations (one each in the United States and Canada) operating at Ku-band send and receive signals to mobile terminals on the forward and return links. These earth stations provide public switched telephone and data network connectivity to the L-band mobile terminals operating anywhere in the coverage area. Mobile satellite ventures (MSV) operates these satellites to provide two-way voice and mobile data services to oil and gas companies, trucking companies (fleet management), as well as recreational and fishing vessels among its many customers. Another major application of these services has been for emergency response, public safety, and homeland security. MSV plans to launch its next generation of mobile satellites in 2009 with significantly enhanced capacity and services including IP-based voice, data, and video mobile broadband services to a handheld mobile device, which would

TABLE 10.1 Primary S-, C-, and Ku-Band Frequency Allocations

Frequency Band (GHz)	Band Designation	Service
2.500–2.535	S	Fixed regional systems, space-to-earth
2.655–2.690		Fixed regional systems, earth-to-space
3.625–4.200	C	Fixed, space-to-earth
5.850–6.425		Fixed, earth-to-space
11.700–12.500	Ku	Fixed, space-to-earth
14.000–14.500		Fixed, earth-to-space

operate transparently with an ancillary terrestrial network (ATN) [19]. MSV approached the FCC with this novel hybrid mobile satellite system architecture and was granted the first ever ATC license to offer MSS services inside the buildings, in urban as well as rural communities [20]. More will be discussed later on this unique MSS concept and architecture.

10.2.4 Frequency Allocations

The frequencies allocated for traditional commercial FSS in popular S-, C-, and Ku-bands [21] are listed in Table 10.1.

At the 1992 World Administrative Radio Conference (WARC) [21,22], the L-band frequencies were made available for mobile communications (Figure 10.4). The L-band allocations have been extended due to a rapid increase in mobile communications traffic. A number of new satellite systems, such as Iridium, Globalstar, and ICO-Global [9], have been implemented to provide personal communications services to small handheld terminals. These systems use LEO between 600 and 800 km (Iridium, Globalstar), or an intermediate (10,000 km) altitude circular orbit (ICO). In addition, a number of goestationary (GEO) systems such as ACeS and Thuraya have been deployed. A detailed discussion of these systems is available in the next chapter. In a significant ruling in 2003, in response to application from the MSV, FCC has allowed the satellite L-band and S-band allocations to be used in the ancillary terrestrial component (ATC) of the network [23] thus allowing further flexibility in the mobile satellite systems.

The 1997 World Radio Conference (WRC-97) adopted Ka-band frequency allocations for geostationary orbit (GSO) and nongeostationary orbit (NGSO) satellite services. Some of these bands require coordination with local multipoint distribution services (LMDS) and/or NGSO MSS feeder links (Figure 10.5). In the United States, the FCC has allocated V-band frequencies for satellite services. In November 1997, the ITU favored the use of 47.2- to 48.2-GHz spectrum for stratospheric platforms. The Ka- and V-band frequency allocations are given in Table 10.2.

10.2.5 Satellite Orbits

Commercial satellite systems were first implemented for operation in GEO orbit 35,700 km above the earth. With GEO satellite systems, a constellation of three satellites is sufficient to provide "approximately global" coverage. In 1990s, satellite systems (led by Iridium) were proposed using LEO satellites (at 700–1800 km altitude) and MEO satellites (at 9000–14,000 km altitude). Figure 10.6 shows the amount of earth's coverage obtained from LEO, MEO, and GEO orbits. The higher the satellite altitude, the fewer satellites required to provide global coverage. The minimum number of satellites required for a given orbital altitude and user elevation angle is shown in Figure 10.7 and the number of orbital planes is shown in Figure 10.8. LEO satellites tend to be smaller and less expensive than MEOs, which in turn are likely to be less expensive than GEO satellites. In terms of space segment implementation, this represents an important system/cost trade-off. For example, the benefit of smaller and less expensive satellites for LEO, as compared to GEO, is offset by the greater number of satellites required.

Another key consideration is the link margin required for the terminal for different satellite orbits. The LEO system offers the advantage of lower satellite power and smaller satellite antennas; however,

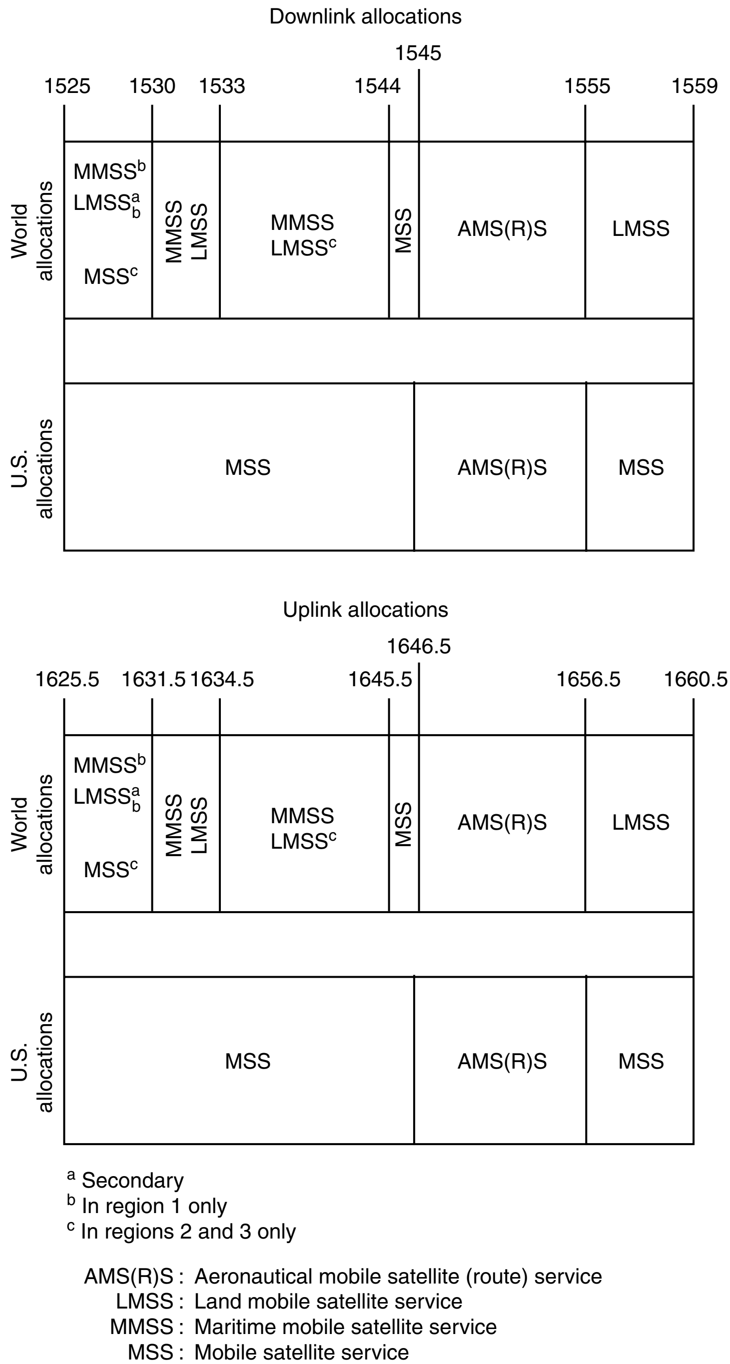


FIGURE 10.4 ARC-92 L-band mobile satellite service allocations.

the user terminals must track the satellites, and an effective handover from satellite to satellite must be designed into the system. To preserve the link margins, the spot size of the beams must be kept small. This requires use of larger satellite antennas, the further out the satellite is placed. For example, required satellite antenna diameters may range from 1 to 12 m for LEO (1 m), MEO (3–4 m), and GEO systems (10–12 m) for closing the link with a similar handheld device.

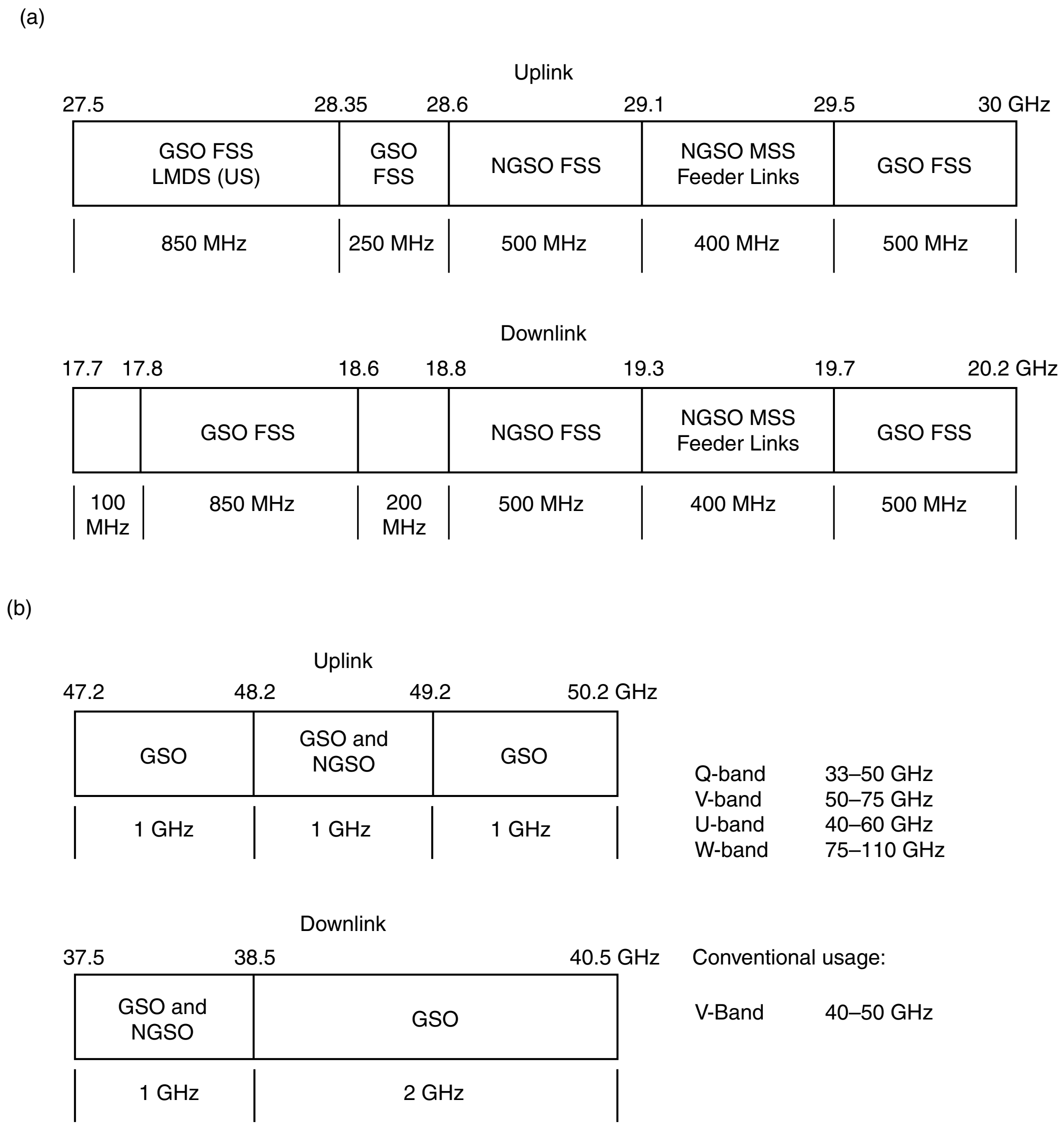


FIGURE 10.5 (a) WARC-97 Ka-band frequency allocations. (b) FCC V-band frequency allocations.

The important characteristics of the three orbital systems are summarized in Table 10.3. System costs, satellite lifetime, and system growth play a significant role in the orbit selection process. On the other hand, round-trip delay, availability, user terminal antenna scanning, and handover are critical to system utility and market acceptance.

10.3 INTELSAT System Example

The evolutionary trends in INTELSAT communications satellites are shown in Figure 10.9. This illustration depicts an evolution from single-beam global coverage to multibeam coverages with frequency reuse. The number of transponders has increased from 2 on INTELSAT I to 50 on INTELSAT VI, with a corresponding increase in satellite EIRP from 11.5 to 30 dBW in the 4-GHz band, and in excess of 50 dBW

TABLE 10.2 Ka- and V-Band Frequency Allocations

Frequency Band (GHz)	Band Designation	Service
17.8-18.6	Ka	GSO FSS, space-to-earth
19.7-20.2	Ka	GSO FSS, space-to-earth
18.8-19.3	Ka	NGSO FSS, space-to-earth
19.3-19.7	Ka	NGSO MSS feeder links, space-to-earth
27.5-28.35	Ka	GSO FSS, earth-to-space (coordination required with LMDS)
28.35-28.6	Ka	GSO FSS, earth-to-space
29.5-30.0	Ka	GSO FSS, earth-to-space
28.6-29.1	Ka	NGSO FSS, earth-to-space
29.1-29.5	Ka	NGSO MSS feeder links, earth-to-space
38.5-40.5	V	GSO, space-to-earth
37.5-38.5	V	GSO and NGSO, space-to-earth
47.2-48.2	V	GSO, earth-to-space
49.2-50.2	V	GSO, earth-to-space
48.2-49.2	V	GSO and NGSO, earth-to-space

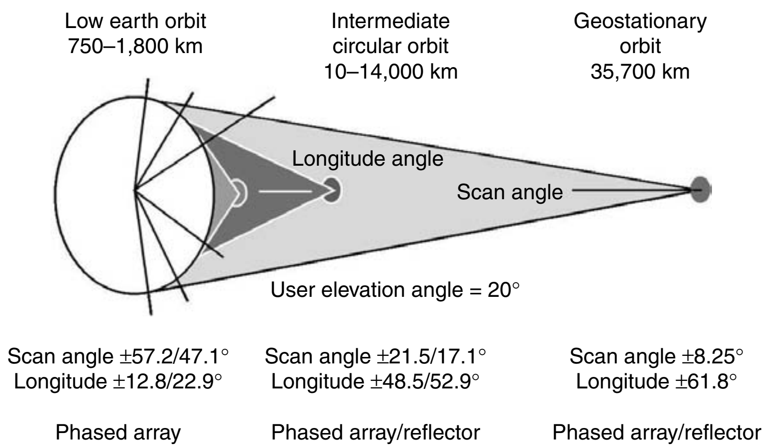


FIGURE 10.6 Relative earth coverage by satellites in LEO, MEO, and GEO orbits.

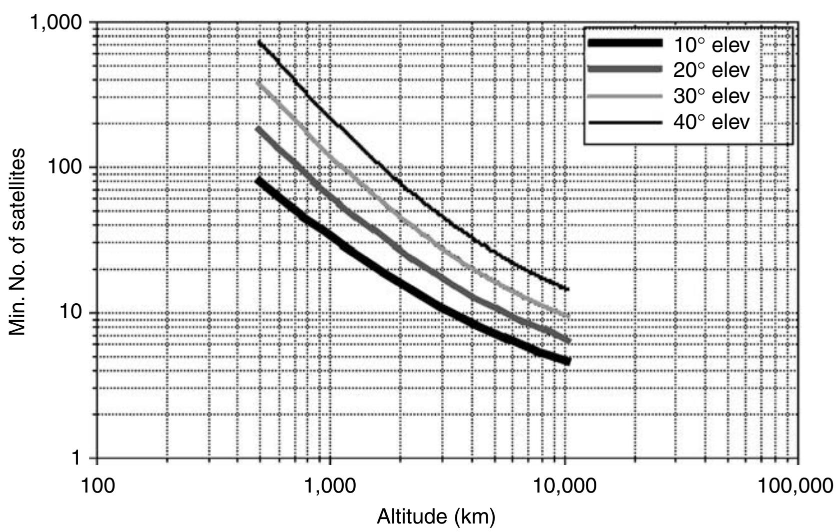


FIGURE 10.7 Number of orbiting satellites versus orbital height.

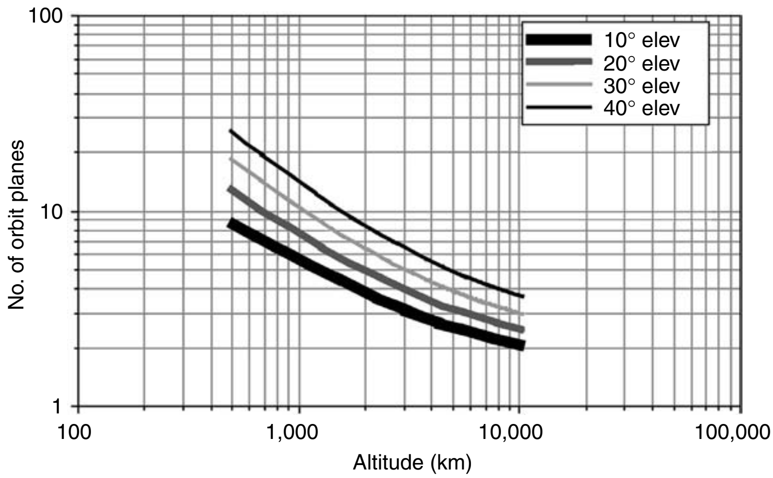


FIGURE 10.8 Number of orbit planes versus orbital height.

TABLE 10.3 LEO, MEO, and GEO Satellite System Characteristics[a]

Characteristic	LEO	MEO	GEO
Space segment cost	Highest	Medium	Lowest
System cost	Highest	Medium	Lowest
Satellite lifetime (years)	5–7	10–12	10–15
Terrestrial gateway cost	Highest	Medium	Lowest
Overall system capacity	Highest	Medium	Lowest
Round-trip time delay	Medium	Medium	Longest
Availability/elevation angles	Poor	Best	Restricted
Operational complexity	Complex	Medium	Simplest[b]
Handover rate	Frequent	Infrequent	None
Building penetration	Limited	Limited	Very limited
Wide area connectivity	Intersatellite links	Good	Cable connectivity
Phased startup	No	Yes	Yes
Development time	Longest	Medium	Shortest
Deployment time	Longest	Medium	Short
Satellite technology	Highest	Medium	Medium

[a] *Source:* A.Williams, et al., "Evolution of personal handheld satellite communications," *Applied Microwave and Wireless*, summer 1996, pp. 72–83.
[b] Recent multibean satellites with multibeam capability may require beam to beam handovers (mobile applications) and are getting more complex to operate.

at Ku-band. During the same timeframe, earth station size has decreased from 30 m (Standard A) to 1.2 m (Standard G) for VSAT data services (Figure 10.10). Several technological innovations have contributed to the increase in the number of active transponders required to satisfy the growing traffic demand [24]. The development of lightweight elliptic function filters resulted in the channelization of allocated frequency spectrum into contiguous transponder channels of 40 and 80 MHz. This channelization provided useful bandwidth of 36 and 72 MHz, respectively, and reduced the number of carriers per transponder and the intermodulation interference generated by the nonlinearity of traveling wave tube amplifiers (TWTAs). For example, using filters and modifying the TWTA redundancy configuration resulted in the provision of twenty 40-MHz transponders in the 6/4-GHz frequency band for INTELSAT-IVA satellites, as compared to 12 for INTELSAT IV.

Traffic capacity was further increased with the introduction of frequency reuse through spatial and polarization isolation. In INTELSAT V, for example, 14/11 GHz (Ku) band was introduced and fourfold

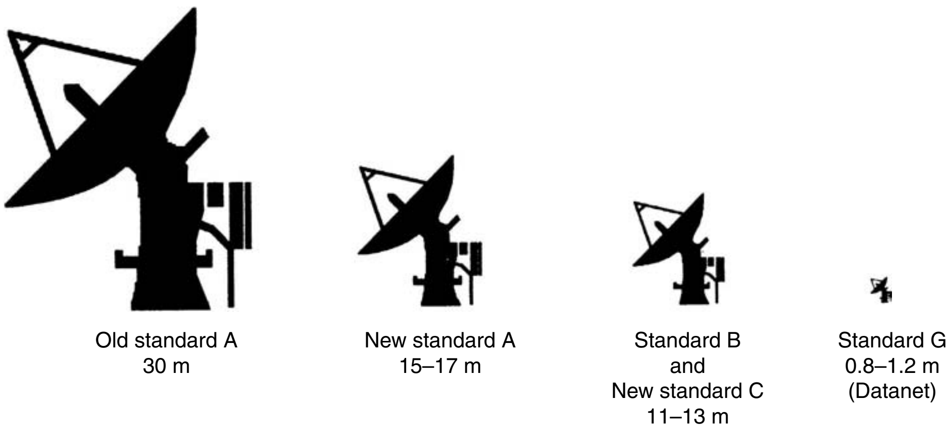
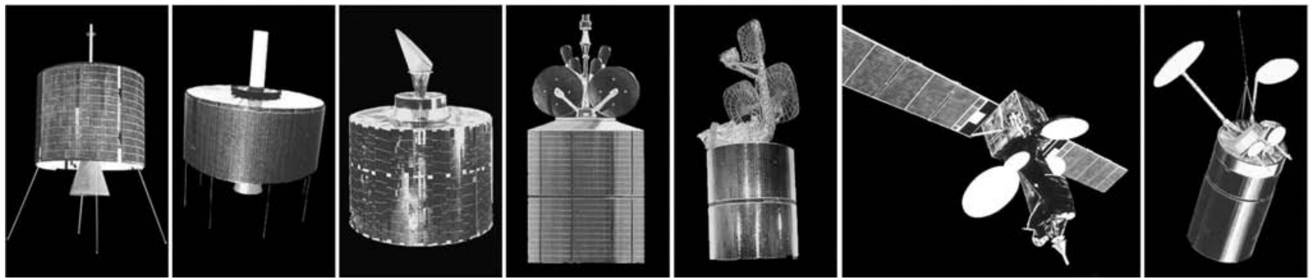


FIGURE 10.9 INTELSAT communications satellite trends. (Courtesy: INTELSAT.)

frequency reuse was achieved at C-band. The use of spatially separated multiple beams also increased antenna gain due to beam shaping, hence increasing the EIRP and gain-to-noise temperature ratio (G/T) for these satellites [25]. This was made possible by significant advances in beamforming and reflector technologies. The increase in satellite G/T and the EIRP enable reductions in the earth terminal antenna size and power for similar link performance. Consequently, earth terminal costs could be reduced and terminals could be located closer to customer premises. Transition also occurred from the analog frequency-division multiple access (FDMA) techniques to time-division multiple access (TDMA) transmission using digitally modulated quadrature phase shift keying (QPSK) signals. The multibeam satellite systems require onboard switch matrices to provide connectivity among the isolated beams. In the INTELSAT V spacecraft, these interconnections were established by using electromechanical static switches that could be changed by ground command. One disadvantage of static interconnections is the inefficient use of satellite transponder capacity when there is not enough traffic to fill the capacity for the selected routing. In the INTELSAT VI spacecraft, satellite utilization efficiency was enhanced by providing cyclic and dynamic interconnection among six isolated beams using redundant microwave switch matrixes (MSMs) [26]. Satellite-switched TDMA (SS-TDMA) operation provided dynamic interconnections, allowing an earth station in one beam to access earth stations in all six beams in a cyclic manner in each TDMA frame [27]. Although INTELSAT VI MSMs were realized using hybrid MIC technology, use of Gallium arsenide (GaAs) monolithic microwave integrated circuit (MMIC) technology was demonstrated for such systems because it offered reproducibility, performance uniformity, and high reliability [28]. Follow-on INTELSAT satellites (INTELSAT-VII, -VIII, -IX) provide additional spot beams for video and Internet services and have also enhanced reliability and life by using GaAs MMIC technology in the payload. Availability of low-cost integrated all-digital variable rate modulators and demodulators has provided additional flexibility in satellite network equipment design and operation. These modem chip sets enable satellite networks to further improve the satellite channel bandwidth utilization efficiency by using variable rate BPSK/QPSK/8PSk and/or 16 QAM with Turbo coding with multiple adaptive or fixed Forward Error Correction (FEC) codes to provide reliable performance for a given satellite channel.

Improvements in quality of service and satellite operational flexibility can be achieved by using onboard regeneration and signal processing, which offer additional link budget advantages and improvements in bit error ratio performance through separation of additive uplink and downlink noise. Use of reconfigurable, narrow, high-gain spot beams with phased-array antennas offers the additional flexibility of dynamic transfer of satellite resources (bandwidth and EIRP). Since these active phased-array antennas require several identical elements of high reliability, MMIC technology makes them feasible [29,30]. These technological developments in microwave and antenna systems, which have helped improve satellite capacity per kilogram of dry mass and reduced cost (Figure 10.11), have positioned satellite systems to launch even higher capacity broadband satellites.

The improved design of INTELSAT satellites has yielded increased capacity and reduced costs for service.



Intelsat designation	I	II	III	IV	IV-A	V	V-A	VI
Year of first launch	1965	1967	1986	1971	1975	1980	1985	1989
Prime contractor	Hughes	Hughes	TRW	Hughes	Hughes	Ford Aerosapce	Ford Aerosapce	Hughes
Width dimensions, m. (undeployed)	0.7	1.4	1.4	2.4	2.4	2.0	2.0	3.6
Height dimensions, m. (undeployed)	0.6	0.7	1.0	5.3	6.8	6.4	6.4	6.4
Launch vehicles	Thor delta	Thor delta	Thor delta	Atlas centaur	Atlas centaur	Atlas centaur Ariane 1,2	Atlas centaur Ariane 1,2	Ariane 4 or NASA STS (shuttle)
Design lifetime, years	1.5	3	5	7	7	7	7	14
Bandwidth, MHz	50	130	300	500	800	2,144	2,250	3,300
Capacity								
Voice circuits	240	240	1,500	4,000	6,000	12,000	15,000	120,000
Television channels	-	-	-	2	2	2	2	3

FIGURE 10.10 INTELSAT earth station size trend.

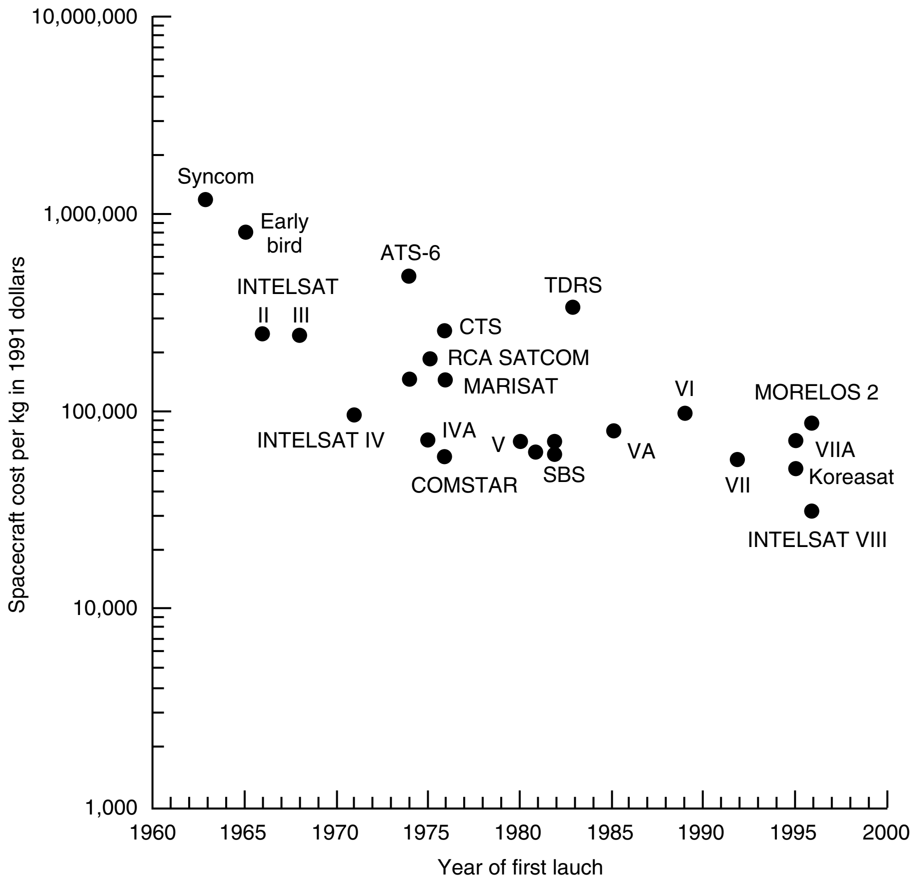


FIGURE 10.11 Communications satellite cost per kilogram.

10.4 Hybrid Satellite Networks

Satellite network are rapidly evolving toward hybrid networks, in which space and terrestrial assets are combined to deliver a complete networking solution to end-customer at most economical prices [31]. As a result, major satellite operators have obtained access to teleport and fiber network facilities through direct acquisition of facilities and/or strategic alliances with teleport operators and network service providers. INTELSAT, the largest satellite operator, for example, owns several teleports including trans-oceanic fiber facilities and has more than 50 points of presence (POPs) enabling provision of satellite-based services to businesses, governments, broadcasters, and cable TV programmers around the world. The hybrid networking concept (Figure 10.12) is largely applied to integration of satellite, fiber, and also cellular transmission capabilities to provide end-to-end solution to the customer including cellular backhaul and two-way connectivity. As broadband third- and fourth-generation (3G and 4G) wireless networks get rolled out, using IEEE 802.16 (WiMax) and broadband CDMA (CDMA2000) standards, satellite hybrid networks would soon be capable of delivering ubiquitous broadband wireless services to fixed and mobile users. It should be emphasized here that Figure 10.12 is a generic and conceptual framework for integration of satellite system(s) in the emerging telecommunications infrastructure, instead of representing a specific system. Figure 10.13 provides a segmentation of satellite systems by different applications, frequency bands, and applicable data rates.

The public switched telephone network (PSTN) evolved toward integrated services data networks (ISDNs) and ATM [32]. ATM offers a suite of protocols that are well suited to handling a mix of voice,

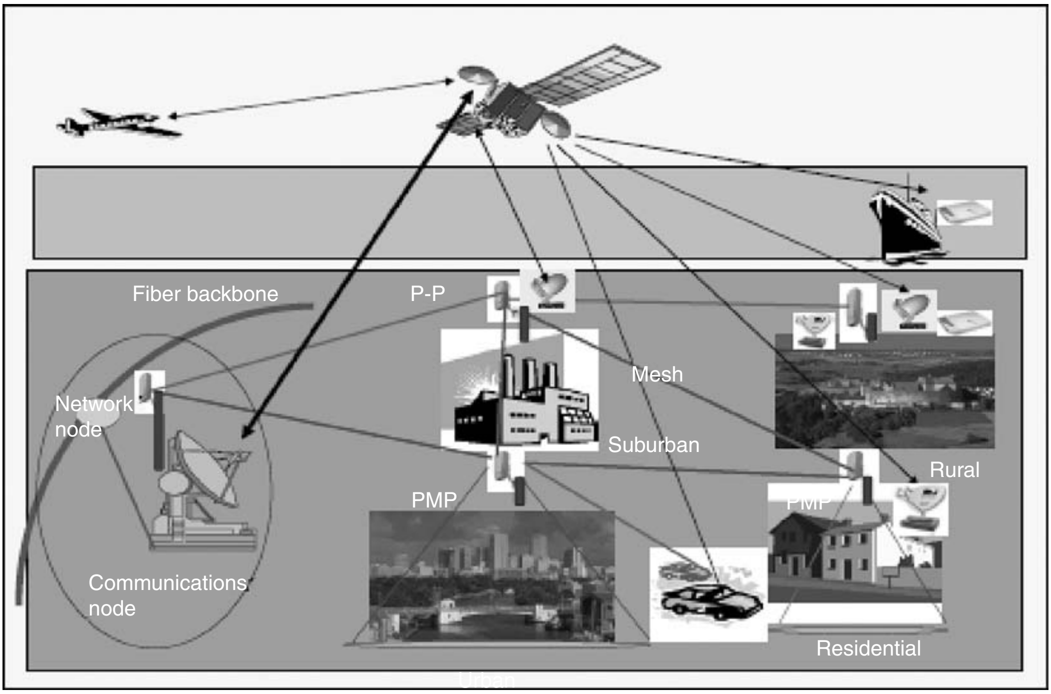


FIGURE 10.12 Hybrid satellite network scenario. (From Ramesh K. Gupta, "Low cost satelite user terminals: lessons from the wireless industry," *22nd AIAA Conference on Satellite Communications Systems*, Monterey, CA, June 2004.)

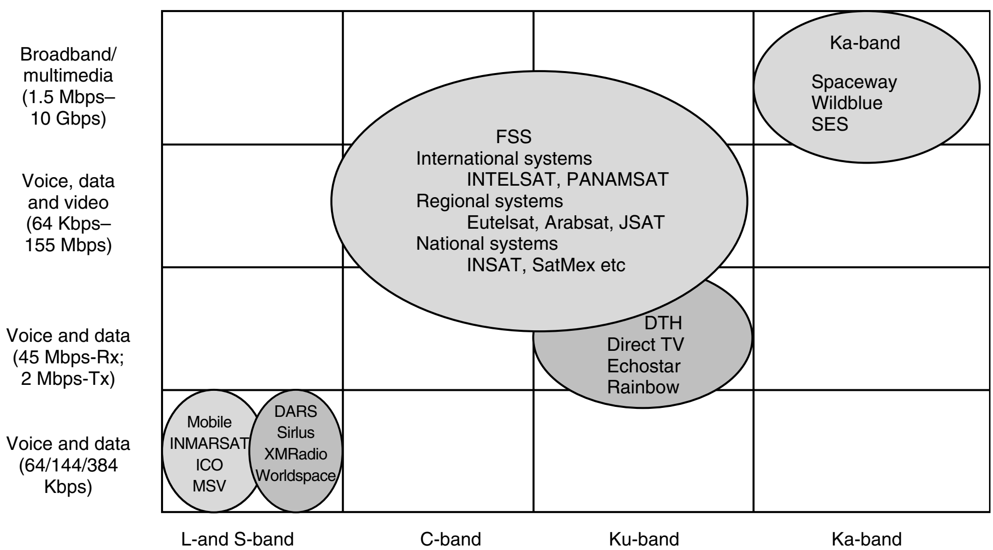


FIGURE 10.13 Satellite service segments by frequency/applications/data rates. (From Ramesh K. Gupta, "Low cost satellite user terminals: lessons from the wireless industry," *22nd AIAA Conference on Satellite Communications Systems*, Monterey, CA, June 2004.)

high-speed data, and video information, making it very attractive for multimedia applications. One of the unique virtues of satellite networks is that the satellite offers a shared bandwidth resource, which is available to many users spread over a large geographical area on earth. This forms the basis for the concept of bandwidth-on-demand, in which terminals communicate with all other terminals (full-mesh connectivity) but use satellite capacity on an as-needed basis. By using multifrequency TDMA (MF-TDMA) to achieve high-efficiency management and flexibility, commercial multiservice networks are implemented using multi-frequency, multirate, TDMA, bandwidth-on-demand mesh networking platforms (COMSAT Laboratories division of ViaSat, HNS, and iDirect). These platforms are capable of providing greater than 2 Mb/s data rates and can provide flexible interfaces (including TCP/IP, ATM, Frame Relay, ISDN) [33]. In MF-TDMA, since each transmitter only transmits a single carrier at a time, with QPSK or 8PSK modulation, it is possible to operate the transmitters close to saturation. An important consideration for seamless integration of satellite and terrestrial networks is the transmission delay and channel fading [34] should not degrade the quality of service (QoS). To overcome interoperability problems, new standards and techniques have been developed for satellite systems including dynamic adaptive coding, data compression, transmission control protocol (TCP) proxy to enable TCP/IP-based applications in the presence of delays associated with geosynchronous satellites, bandwidth-on-demand, and traffic management [33].

Telecommunications Industry Association (TIA) adopted IP over Satellites (IPoS) [35] standards in 2003. This standardization has enabled delivery of "always on" broadband IP services commercially to residential and enterprise markets using, for example, HNS platform—DIRECWAY. Availability of IP Multicast services such as audio/video streaming, distance learning based on this standard has accelerated the adoption of satellite hybrid services. Using this standard, a number of remote terminals can be set up in a star network, so that they communicate bi-directionally with a hub. Space segment typically uses "bent-pipe" satellite transponders and commercial Ku-band spectrum for FSS. The hub segment consisting of a satellite gateway communicates with a large number of user terminals, aggregates traffic, provides connection to public and/or private data networks, converts traffic from one protocol to other, and provides overall network management and configuration of the satellite access network. In the gateway, Satellite-Independent service Access Point (SI-SAP) interface provides separation between satellite dependent and satellite-independent functions. The remote terminal may be PC hosted, where the terminal acts as a peripheral typically using the USB port, or self-hosted independent terminals. The return channel may be either provided directly through the satellite channels or may be provided using a terrestrial return, including a dial-up phone connection. For high-speed links between Internet backbone and a customer POP, links are made using Direct Video Broadcast (DVB) satellite standard for shared data rates up to 45 Mbps [36] (e.g., PANAMSAT- SPOT-byte, New Skies-IP Sys Premier). Ku (or Ka-band) Return Channel by Satellite (DVB-RCS) standard has also been developed and is presently in use for hybrid satellite networks.

Ku-band and C-band transponders have been used by most FSS operators to offer hybrid satellite services. Ku-band antenna sizes are typically 1.8 m whereas C-band antenna sizes tend to be 2.4–3.7 m. For a hubless mesh connectivity, the uplink RF power at Ku- and C-band can be up to 16 or 40 W, respectively. For a system configuration with a large hub, the transmit power is reduced to a more reasonable 2–5 W making the terminals lot less expensive. These systems operate in a star connectivity, thus requiring double-hop operation for communication between user terminals. To fully exploit the bandwidth-on-demand capability by using MF-TDMA with frequency hopping over the full satellite band, availability of low cost broadband ground RF terminal equipment is critical. Several cost reduction issues related to the terminal RF equipment include low-cost, high power integrated solid-state power amplifiers (SSPA), a single inter facility link (IFL) connection between the RF outdoor unit (ODU) and the indoor unit (IDU), and multiplexed monitoring and control (M& C) and DC power over the IFL. Significant cost reductions continue to be achieved for the SSPA's using integrated GaAs MMIC PA modules.

Examples of hybrid satellite networks, where satellite and terrestrial resources are combined to provide distance education, include WIDE (widely integratted distribution environments) project [37] in Japan and African Virtual University Project [38]. In the WIDE project [named as School on Internet (SOI)—Asia] and supported by several ministries of the Japanese government, 12 Asian countries are linked together using a JSAT satellite and terrestrial links. Satellite is used to provide multicast and (video/audio) streaming of lectures, whereas return link is provided using satellite channels or existing terrestrial infrastructure (128 kbps–1.5 Mbps) for real-time interactive sessions. For the AVU project, video and data IP traffic is broadcast over a wide area using Intelsat-8 Ku-band transponders with DBV over satellite (DVB-S) standard and a reverse asymmetric Internet access. During the off-peak hours, the network is used to "push" learning materials asynchronously to the learning centers. These C- and Ku-band networks demonstrate technologies that can be used with broadband Ka-band satellite networks capable of supporting significantly higher data rates because of available bandwidth.

10.5 Broadband Ka-Band Satellite Networks

NASA's advanced communications technology satellite (ACTS), which was launched in September 1993, demonstrated new system concepts, including the use of Ka-band spectrum, narrow spot beams (beam widths between 0.25° and 0.75°), independently steerable hopping beams for up- and downlinks, a wide dynamic range MSM, and onboard baseband processor. In addition to these technology developments, several experiments and demonstrations have been performed using ACTS system. These include demonstration of ISDN, ATM, and IP using Ka- band VSATs, high-definition video broadcasts, health care and long-distance education, and several propagation experiments.

The success of the ACTS program, together with the development of key onboard technologies, a better understanding of high-frequency atmospheric effects, and the availability of higher frequency spectrum has resulted in proposals for a number of Ka- and V-band satellite systems [8,9]. Of the 14 proposed Ka-band systems in the United States, only two—Spaceway (Hughes Networks) [39] and Wildblue [40] have been implemented. None of the proposed V-band systems has gone beyond the filings and FCC approval stage. Existing satellite operators have opted to reduce their Ka-band market risk by augmenting their new generation satellites with Ku-/Ka-band transponders. Examples of satellite systems with Ka-band payload include Astra 1K (SES Global–Europe), Superbird (Japan and Asia), Anik F3 (Telesat—Canada), NSS-6 (New Skies), and Intelsat 8 (Intelsat—Asia/Africa) [41]. Ka-band transponders in SES satellites are being used with Ku-band DBS services with Ka-band return link. In Asia, higher gain Ka-band spot beams (e.g., NSS-6) are being used for satellite news gathering (SNG) services. In the United States, Ka-band satellites are used to provide high definition TV (HDTV) service because of limited availability of Ku-band capacity for broadband TV transmissions. The first two Spaceway satellites were modified to provide Ka-band HDTV services by DIRECTV. There is also a slow penetration of Ka-band enterprise satellite networks in the marketplace as Ka-band satellite services become available.

Wildblue and Spaceway (F3) are specifically targeting Ka-band consumers and SOHO markets [39,40]. Wildblue started offering Ka-band Internet services to consumers using Anik-F3 satellite. Wildblue-1 provides a bent-pipe satellite system for low-cost broadband satellite services to consumers. In contrast, the next generation Spaceway (F3) satellite platform exploits advanced multi-spot beam antenna and onboard data processing technologies to provide full-mesh (single hop) interconnectivity between the user terminals. Use of narrow spot beams provides ability to reuse the frequency 8–10 times with onboard processor demodulating the bits and providing onboard packet level routing between the uplink and downlink beams and then remodulating the signals for transmission to ground. Using these advanced onboard technologies and satellite network architecture, Spaceway is capable of delivering upload speeds from 512 kbps to 16 Mbps and download speeds as high as 30 Mbps.

10.5.1 Key Technologies for Broadband Satellite Networks

A number of advanced satellite technologies have contributed to the evolution of broadband FSS networks. Owing to successful technology demonstration provided at Ka-band by the ACTS program, technologies including hopping beams, onboard demodulation/remodulation, onboard FEC decoding/coding, baseband switching (BBS), and adaptive fade control by FEC coding and rate reduction will continue to be used for any future Ka (or V-band) systems. However, some of the proposed broadband satellite systems are vastly more complex than ACTS (e.g., a system capacity of 220 Mbps in ACTS versus 16 Gbps in the proposed systems) and also employ new processing/switching technologies such as multicarrier demultiplexer/demodulators for several thousand carriers, fast packet switching (FPS), and intersatellite links (ISLs). Three of the key technology areas that significantly influence payload and terminal design—onboard processing, multibeam antennas, and propagation effects—are discussed in the subsections that follow.

10.5.1.1 Onboard Processing

Future broadband multimedia satellite systems are likely to employ onboard processing/switching [8], although some will use "bent-pipe" or SS-TDMA operation onboard the satellites. A majority of the processing systems will employ FPS, which is also referred to as cell switching, packet switching, asynchronous transfer mode (ATM) switching, and packet-by-packet routing in various FCC filings. Some of the proposed systems are currently undecided regarding their BBS mechanisms and will probably use either FPS or circuit switching. Along with onboard BBS, the processing satellites will most likely employ digital multicarrier de-multiplexing and demodulation (MCDD). Onboard BBS allows optimized transmission link design based on user traffic volume, and flexible interconnection of all users in the network at the satellite. ISLs will provide user-to-user connection in many of these systems without assistance from ground stations.

Gallium arsenide (GaAs) MMIC technology has been used successfully to develop MSM arrays for SS-TDMA operation and RF demodulator/remodulator hardware. Development of low-power, application-specific integrated circuits (ASICs) with high integration densities and radiation tolerance is continuing and is critical to the realization of relatively complex onboard processing and control functions.

10.5.1.2 Multibeam Antennas

The design of the satellite antennas depends on the beam definition, which in turn is a function of system capacity and projected traffic patterns. Several systems require a large number of fixed, narrow spot beams covering the whole service area and designed to deliver a high satellite EIRP (>60 dBW) to user terminals. A single reflector with a large number of feeds may provide such coverage. However, if scanning loss is excessive due to the large number of beams scanned in one direction, multiple reflectors may be required. The coverage area is divided into a number of smaller areas, with each reflector boresight at the center of the corresponding area. Similarly, single or multiple phased arrays may be used. The phased array may have lower scan loss and, with the flexibility of digital beamformers, a single phased array can handle a large number of beams with no added complexity. If the system design calls for a small number of hopping beams instead of a large number of fixed beams, the phased array solution becomes more attractive due to the flexibility and reliability of the beamformer versus the switching arrangement in a focal-region-fed reflector antenna. For a small number of beams at a time, the microwave beamformer becomes a viable alternative and the choice between the microwave and the digital beamformers becomes a payload system issue.

At Ka-band, the manufacturing tolerance of the array feed elements and the surface tolerance of the reflectors play an important role in overall antenna performance. Waveguide-based elements are well developed at Ka-band, but lighter weight, lower profile printed circuit elements may need further development for space applications. For a large number of beams and large number of frequency reuses, co- and cross-polarization interference becomes a major issue that imposes severe restrictions on the antenna side lobe and cross-polarization isolations. Although the receive antenna may benefit from statistical averaging based on the users distribution, the transmit antenna must satisfy a certain envelope in order to meet the required interference specifications.

10.5.1.3 Propagation Effects

Line-of-sight rain attenuation and atmospheric propagation effects are not significant at L-, S-, and C-bands. At high elevation angles, the communications between satellites and terminals at L- and S-bands is very reliable. In the mobile environment, however, multipath effects and signal blockages by buildings cause signal fades, which require cooperative users willing to change their location. In addition, the links must be designed at worst-case user elevation and for operation at the beam edge. Because of these considerations, 10- to 15-dB link margin is designed into the systems.

In comparison, the troposphere can produce significant signal impairments at the Ku-, Ka-, and V-band frequencies, especially at lower elevation angles, thus limiting system availability and performance [12,34]. Most systems are expected to operate at elevation angles above about 20° . Tropospheric radio wave propagation factors that influence satellite links include gaseous absorption, cloud attenuation, melting layer attenuation, rain attenuation, rain and ice depolarization, and tropospheric scintillation. Gaseous absorption and cloud attenuation determine the clear-sky performance of the system. Clouds are present for a large fraction of an average year, and gaseous absorption varies with the temperature and relative humidity. Rain attenuation—and to some extent melting layer attenuation—determine the availability of the system. Typical rain time is on the order of 5–10% of an average year. Depolarization produced by rain and ice particles must be factored into the co-channel interference budgets in frequency reuse systems. Signal amplitude fluctuations due to refractive index inhomogenieties (known as *tropospheric scintillation*) are most noticeable under hot, humid conditions at low elevation angles. Scintillation effects must be taken into account in establishing clear-sky margins and in designing uplink power control systems. Figure 10.14 shows the combined clear sky attenuation distribution at Ka- and V-band frequencies for a site in the Washington, DC, area at an elevation angle of 20°. Figure 10.15 shows the rain attenuation distribution for the same site. It can be seen that the required clear-sky margin can be several dBs at V-band, especially at 50 GHz, due to the elevated oxygen absorption. Figure 10.15 indicates that, to achieve reasonable availabilities, rain fade mitigation must be an integral part of the system design. Rain fade mitigation can be accomplished through power control, diversity, adaptive coding, and data rate reduction.

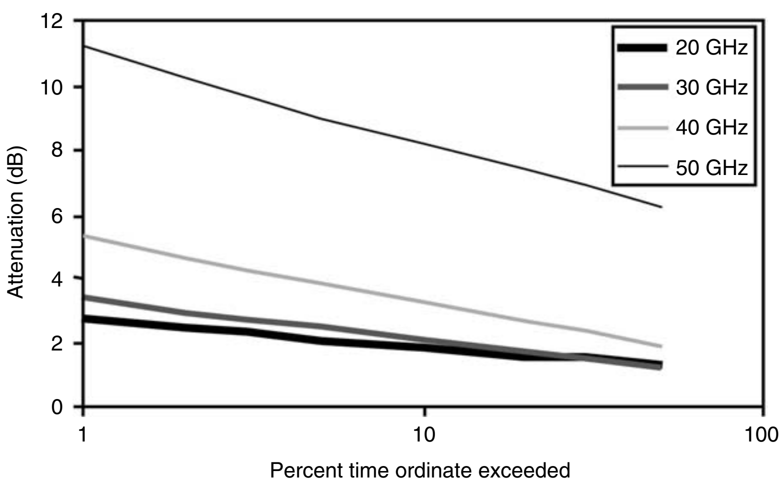


FIGURE 10.14 Probability distribution of clear-sky attenuation at 20° elevation, Washington, DC.

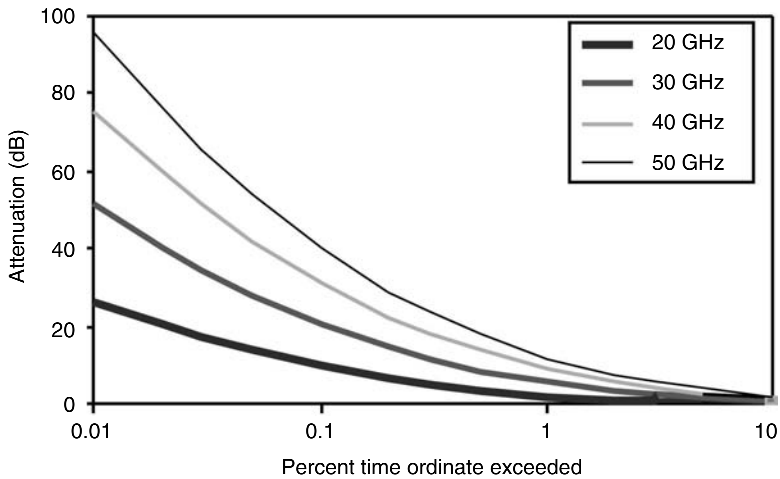


FIGURE 10.15 Probability distribution of rain attenuation at 20° elevation, Washington, DC.

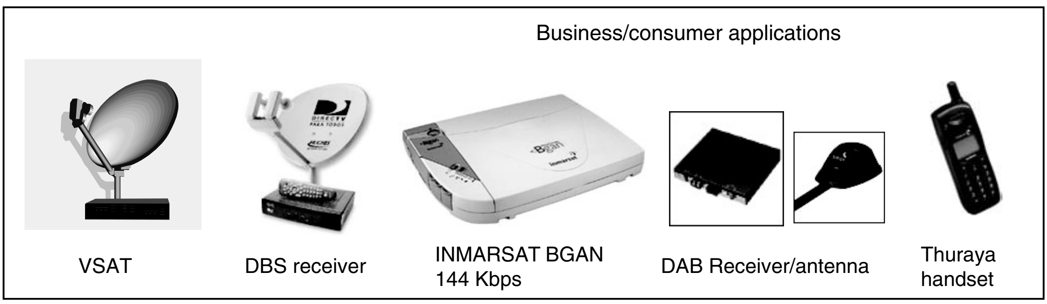


FIGURE 10.16 Trends in satellite terminals for consumer/business applications. (From Ramesh K. Gupta, "Low cost satellite user terminals: lessons from the wireless industry," *22nd AIAA Conference on Satellite Communications Systems*, Monterey, CA, June 2004.)

10.6 User Terminals

Satellite systems and hybrid satellite networks are expected to play a major and continuing role in the distribution of IP data, video, and multimedia content to fixed and mobile users. An important barrier to growth of satellite networks has been the availability of low cost user terminals. This is in part due to the fact that majority of satellite systems are designed with unique system requirements (including frequency bands and modes of operation), thus hindering volume terminal production and not achieving economies of scale. Industry associations like global VSAT forum (GVF) [42] have been formed and have promoted some standardization of antenna and RF equipment to reduce the RF terminal costs. Satellite-based Direct Broadcast video and radio services have also stimulated the development of low cost satellite receive terminals. For example, DBS services achieved an impressive market penetration for video services because of a low cost (<$300) terminal, which could be offered as part of bundled service offering. Similarly, for satellite radio services satellite operators (Sirius, XMRadio, and Worldspace) developed a low cost receiver ($100–$200) before launching their service. As a result, the number of in-car receivers is growing with many automobile manufacturers offering the satellite radio receiver as standard equipment. The size of dual- or triple-mode satellite phones used for narrow band voice and data services has significantly reduced from a "brick" size to a small handheld device. Figure 10.16 shows this trend in the satellite terminal size reduction for different corporate or consumer applications. The prices for these user terminals remain between $200 and $2000 ranges. Further volume driven price reductions are possible through use of low-cost RF chip sets and power amplifiers and higher level of integration in the baseband chip-set with flexibility to support QPSK, 8-PSK, 16-QAM, and even 64-QAM modulation schemes [30].

For the Ka-band satellite systems, a typical user terminal operates at an uplink bit rate between 128 kb/s and 2 Mb/s and a downlink rate up to 10 Mb/s. These terminals employ small-aperture antennas with diameters of 50–66 cm as well as a SSPA of 1–10 W. Uplink power control is required for all of these terminals. Terminal antennas for LEO and MEO systems must provide tracking capability to perform handovers every few minutes. All RF components (SSPA, LNA, and up/down-converters) are integrated into a small ODU and antenna system. Development of low-cost antennas and low-cost, low-power consuming high-frequency RF integrated circuits (RFICs) is critical to substantially reducing the cost of the terminals to the targets of 500–$2000.

10.7 Mobile Satellite Systems (MSS) with ATC

Hybrid satellite-terrestrial networks offer the advantages of combining unique and most powerful features of each of the access networks with ability to provide seamless value-added services to the end-users. The next generation mobile satellite system with ancillary terrestrial component (ATC) to be operated by MSV at L-band is an example of such a hybrid satellite network [20]. The MSV next generation system consists of two integrated networks: a space based network (SBN) and an ancillary terrestrial network (ATN) consisting of several ATCs. The network will be designed so that a single handheld cellular/PCS type device is capable of providing seamless and ubiquitous two-way high-speed data services including voice. In this system, the terrestrial component ensures availability of broadband services at affordable prices in major urban areas with higher population densities, whereas satellite component provides wide-area coverage in low-density rural population areas. The inherent disaster tolerance of satellite networks offers the added advantages of service availability and capacity re-allocation, in case terrestrial infrastructure is disabled by disasters like hurricanes (e.g., Katrina in 2005) and earthquakes and so on. This hybrid architecture uniquely addresses the public safety needs while also offering affordable communications services to the public in remote areas. In response to MSV's 2001 application, the FCC ruled in 2003 that the public interest would be served in authorizing MSS operators to use ATC. L- and S-band spectrum allocated for MSS in North America was authorized for hybrid mobile services with ATC. MSV plans to offer these services using L-band spectrum while Terrestar Networks and ICO Global plan to offer these services at S-band.

A simplified diagram of MSV's next generation L-band MSS/ATC hybrid satellite/ATC network is illustrated is Figure 10.17. Space segment in this illustration consists of two L-band geostationary satellites. These satellites have sufficient aggregate EIRP (AEIRP) and G/T to enable direct link with a small cellular/PCS like handheld device with a significant link margin on the forward and return links. High AEIRP and G/T are achieved by using a very large aperature (22 m) deployable reflector and also by forming a large number of narrow spot beam to provide Continental U.S. (Conus) including Alaska and Hawaii, Canada and Mexico coverage. Communications link can be established with margin by using a single satellite. The second satellite provides diversity combining of the user signals on the return path to achieve additional gain for potentially faded mobile channels (multipath effects and shadows) and also acts as an in-orbit spare. Multiple satellite gateways are connected to the satellite via Ku-band feeder links and serve as additional communications nodes in the terrestrial network. These L-band satellite and terrestrial gateways are connected into the same core switching (PSTA/PDN) network resulting in seamless and transparent operation between the satellite and terrestrial networks. One of the significant features of this hybrid satellite/ terrestrial architecture is that both the satellite and terrestrial network will support the same air interface standard [e.g., GSM, W-CDMA (CDMA2000), WiMax (802.16), and others]. Satellite adaptation of these terrestrial standards will take place at the satellite gateways to compensate for satellite delay, motion (Doppler) and so forth, thus making the satellite-based network design practically independent of air interfaces.

Terrestrial component of this hybrid network ensures scalability of the overall network, whereas satellite network provides wide-area coverage and also flexibility of capacity allocation as needed. Spectrum usage efficiency is achieved by using MSV's patented frequency reuse approach between the satellite and terrestrial segments of the overall hybrid network, through capacity allocations made from the Network

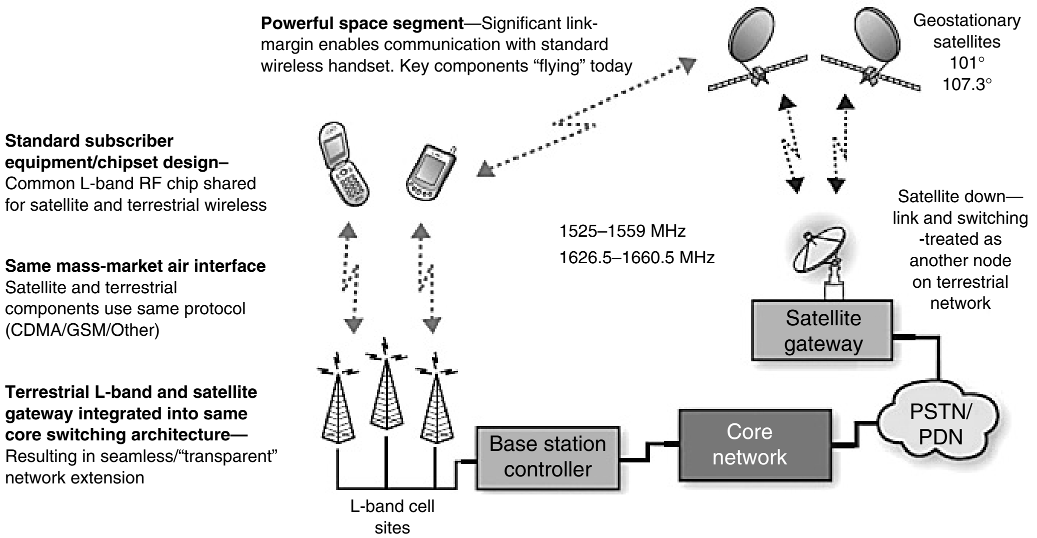


FIGURE 10.17 Hybrid terrestrial/satellite wireless network. (From P. Karabinis and S. Dutta, "Recent Advances that may revitalize Mobile Satellite Systems" MSV; www.msvlp.com.)

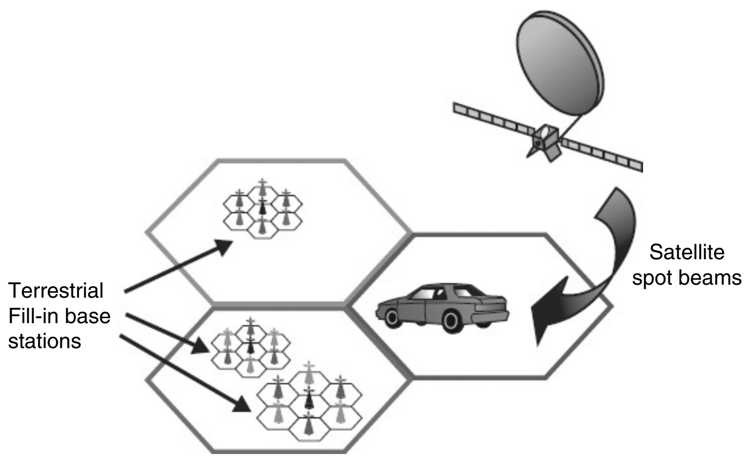


FIGURE 10.18 Satellite/ATC spectrum reuse concept. (From P. Karabinis and S. Dutta, "Recent Advances that may revitalize Mobile Satellite Systems" MSV; www.msvlp.com.)

Operations Center. The satellite beams can be formed and shaped by using a ground based beamforming (GBBF), where the beam signals are digitized, divided in to component beams so that by selecting appropriate amplitude and phase weights beam shapes can be optimized for maximum gain over the desired areas while suppressing ATC signals induced interference in the satellite network [43]. As apparent from Figure 10.18, the satellite cells have a large coverage area (150–300 km across) so that terrestrial cells (1–3 km across) are nested inside the satellite spot beams. Network management techniques will allow seamless transition between the satellite and terrestrial modes. This hybrid network is significantly different from Thuraya, Iridum, and Globalstar mobile satellite networks, where a dual-mode handset is used to provide "interworking" between the satellite and terrestrial networks operating over different frequency bands. That results in a bulkier and a more expensive user device. MSS/ATC hybrid satellite system uses the same frequency band and air interface thus providing ubiquitous communications service offerings. Some of the value-added broadband service offerings for these networks are likely to include public safety and homeland security, fleet management, return link for interactive television with DBS services, and rural market services.

10.8 Summary

The broadband telecommunications services are growing because of increased demand for IP data, voice, and video traffic fueled by growth in the fixed and mobile Internet. Multimedia services and hybrid satellite networks are now beginning to emerge to provide end-to-end value added services to mobile and fixed users. Satellites also provide "instant infrastructure" after system launch and are therefore viewed as a cost-effective solution for providing wide area coverage for developing countries. In the developed world, satellites could provide an effective network node in the telecommunications networks to provide broadband video and IP data services to homes and businesses. Disaster recovery and emergency communications for public safety remains an important application for existing and emerging satellite systems. Hybrid satellite networks integrated with terrestrial infrastructure are already playing a vital role in distance learning, telemedicine, and providing high-speed solutions to enterprises and homes. Innovations in MSS and ATC hybrid networks are likely to result in ubiquitous and seamless operation of wideband wireless services and unlock the potential of MSS networks providing wireless communication anywhere, anytime at affordable prices.

Acknowledgments

The author gratefully acknowledges the contributions of many of his past and present colleagues in the industry, in particular at COMSAT Laboratories and MSV. Author's individual and joint work with many of them has shaped the contents of this chapter. The author would like to dedicate this chapter to the memory of professor K. C. Gupta, a teacher, mentor and past president of IEEE MTT-S, who passed away on February 7, 2007.

The views expressed in this chapter are those of the author and do not represent the opinion of MSV, its employees, or any of its affiliates.

References

1. J. Evans, "Network interoperability meets multimedia," *Satellite Communications*, February 2000, pp. 30–36.
2. P. Chitre and R. Gupta, "Satellite systems for multimedia and internet traffic," *2001 MTT-S International Microwave Symposium Digest*, Phoenix, AZ, May 2001, pp. 1129–1132.
3. "IP in the Air" Cables and Satellite Europe, January 2001.
4. "Multimedia and Satellites—Buying the Whole Package," Via Satellite, February 2001, pp. 26–34.
5. http://www.Intelsat.com
6. B. Pontano, "Satellite communications: services, systems, and technologies", *1998 IEEE MTT-S International Microwave Symposium Digest*, June 1998, pp. 1–4.
7. J. Evans, "Satellite and personal communications," *15th AIAA International Communications Satellite Systems Conference*, San Diego, CA, February/March 1994, pp. 1013–1024.
8. J. Evans, "Satellites systems for personal communications," *IEEE Antenna and Propagation Magazine*, Vol. 39, June 1997, pp. 7–10.
9. A. Williams, R. Gupta, and A. Zaghloul, "Evolution of personal handheld satellite communications," *Applied Microwave and Wireless*, Summer 1996, pp. 72–83.
10. S. Lee et al., "The wireless broadband system for broadband wireless internet services," *IEEE Communications Magazine*, July 2006, pp. 106–113.
11. J. Evans, "The US filings for multimedia satellites: a review," *International Journal of Satellite Communications*, Vol. 18, 2000, pp. 121–160.
12. J. Evans and A. Dissanayake, "The prospectus for commercial satellite services at Q- and V-band," *Space Communications*, Vol. 15, 1998, pp. 1–19.

13. S. Bennett and D. Braverman, "INTELSAT VI—A continuing evolution," *Proceedings of IEEE*, Vol. 72(11), November 1984, pp. 1457–1468.
14. N. Wong, "INTELSAT VI—A 50-channel communication satellite with SS-TDMA," *Satellite Communications Conference*, Ottawa, Canada, June 1983, pp. 20.1.1–20.1.9.
15. I. Koffman, "Hybrid Satellite-Terrestrial Networks, A Novel Concept for Interactive Digital Television," Runcom Technologies Ltd white paper, www.runcom.com.
16. B. Gallagher, ed., *Never Beyond Reach—The World of Mobile Satellite Communications*, International Maritime Satellite Organization, 1989.
17. Fourth Generation Satellite and BGAN Service www.Inmarsat.com
18. D. Whalen and G. Churan, "The American Mobile Satellite Corporation Space Segment," *14th AIAA Conference on Satellite Communications*, Washington, DC, March 1992, pp. 394–404.
19. P. Karabinis and S. Dutta, "Recent Advances that may revitalize Mobile Satellite Systems" MSV (www.msvlp.com).
20. "An ATC Primer: The Future of Communications," Mobile Satellite Ventures white paper, www.msvlp.com.
21. E. Reinhart and R. Taylor, "Mobile communications and space communications," *IEEE Spectrum*, February 1992, pp. 27–29.
22. J. Evans, "Satellite and personal communications," *15th AIAA International Communications Satellite Systems Conference*, San Diego, CA, February–March 1994, pp. 1013–1024.
23. Report and Order and Notice of Proposed Rulemaking, FCC-03-15, *Flexibility of Communications by Mobile Satellite Service Providers in the 2 GHz Band, the L-Band, and the 1.6/2.4 GHz Bands*, IB Docket No. 01-185, Adopted January 29, 2003, Released: February 10, 2003.
24. R. Gupta and F. Assal, "Transponder RF technologies using MMICs for communications satellites," *14th AIAA Conference on Satellite Communications*, March 1992, pp. 446–454.
25. R. Sorbello et al., "A Ku-band multibeam active phased array for satellite communications," *14th AIAA Conference on Satellite Communications*, Washington, DC, March 1992, pp. 1256–1260.
26. F. Assal, R. Gupta, K. Betaharon, A. Zaghloul, and J. Apple, "A wideband satellite microwave switch matrix for SS-TDMA communications," *IEEE Journal on Selected Areas in Communications*, Vol. SAC-1(1), January 1983, pp. 223–231.
27. R. Gupta, J. Narayanan, A. Nakamura, F. Assal, and B. Gibson, "INTELSAT VI On-board SS-TDMA subsystem design and performance," *COMSAT Technical Review*, Vol. 21(1), Spring 1991, pp. 191–225.
28. R. Gupta, F. Assal, and T. Hampsch, "A microwave switch matrix using MMICs for satellite applications," *IEEE MTT-S International Microwave Symposium*, Dallas, TX, May 1990, *Digest*, pp. 885–888.
29. R. Gupta et al., "Beam-forming matrix design using MMICs for a multibeam phased-array antenna," *IEEE GaAs IC Symposium*, Monterey, CA, October 1991, pp. 41–44.
30. Ramesh K. Gupta, "Low cost satelite user terminals: lessons from the wireless industry," *22nd AIAA Conference on Satellite Communications Systems*, Monterey, CA, June 2004.
31. R. Gupta, "MMIC insertion in communications satellite payloads," Workshop J: GaAs MMIC system insertion and multifunction chip design, *IEEE MTT-S International Microwave Symposium*, Boston, MA, June 1991.
32. D. Chitre, D. Gokhale, T. Henderson, J. Lunsford, and N. Matthews, "Asynchronous transfer mode (ATM) operation via satellites: issues, challenges, and resolutions," *International Journal of Satellite Communications*, Vol. 12, 1994, pp. 211–222.
33. D. Chitre, "Interoperability of satellite and terrestrial networks," *IEEE Sarnoff Symposium*, Trenton, NJ, April 2000.
34. A. Dissanayake, J. Allnutt, and F. Haidra, "A prediction model that combines rain attenuation and other propagation impairments along the earth-satellite path," *IEEE Transactions on Antennas and Propagation*, Vol. 45, No. 10, October 1997.
35. IP Over Satellites (IPoS)—The Standard for Broadband Over Satellites—HNS-30740, Hughes Networks Systems, Inc. (www.hns.com).
36. http://www.ipsys.newskies/ipsys.
37. "About SOI-Asia Project," http://www.soi.wide.ad.jp/soi-asia/about/index.html.
38. The African Virtual University (AVU), www.avu.org/avu_tech.asp.

39. "Different Business Models for Different Markets" Viasatellite.com, http: //www.satellitetoday.com/satellite/viasatellite/current/web_exclusive2.htm.
40. www.wildblue.com.
41. www.spaceway.com/HUGHES.
42. Global VSAT Forum, www.gvf.org.
43. P. Karabinis, S. Dutta, and W. Chapman, "Interference Potential to MSS due to Terrestrial Reuse of Satellite Band Frequencies," Mobile Satellite Ventures (www.msvlp.com).

11

Satellite-Based Cellular Communications

Nils V. Jespersen
The Aerospace Corporation

11.1 Introduction

Communication satellite systems designed to serve the mobile user community have long held the promise of extending familiar handheld cellular communication to anywhere a traveler might find himself. One impetus to the fulfillment of this dream has been the success of the Inmarsat system of communication satellites. Founded in 1979 as an international consortium of signatories, Inmarsat provides worldwide communication services to portable and transportable terminals, thereby meeting one of its mandates by enhancing safety on the high seas and other remote areas. Although it does not support handheld, cellular-like operations (due to limitations in the satellite design), the current Inmarsat system is a successful business. Clearly, the next logical step would be to enhance the capabilities of the space segment to provide cell phone utility with even greater ubiquity than available terrestrially. What we would then accomplish is to, essentially, raise the familiar cellular base station several hundreds of kilometers high and, thereby, extend the coverage many times over (see Figure 11.1).

Many elements comprise the optimum solution to a satellite cellular system, not the least of which is the business aspect. The best technical solution does not necessarily result in a successful overall solution. A significant market consolidation is in process at the time of this writing.

In this discussion, we will focus on the systems that are intended to provide voice and/or data services similar to terrestrial cellular communications. As such, we will not be discussing systems intended to provide low-rate messaging and asset tracking services (the so-called little-LEO systems such as Orbcomm and LEO-One). Of the many mobile satellite systems proposed in the late 1990s to provide cellular-like

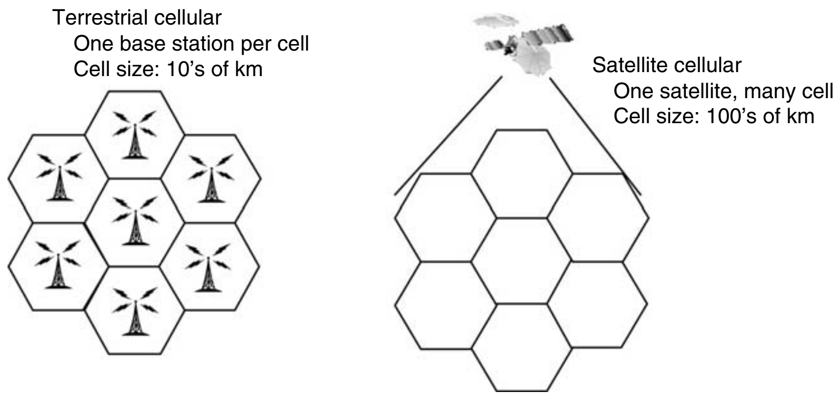


FIGURE 11.1 Comparison: terrestrial versus satellite cellular coverage.

service, only a handful remain as viable. We will examine the salient characteristics of satellite cellular system design and, then, consider some specific system examples.

11.2 Driving Factors

How a particular cellular satellite system is configured depends on a variety of factors combined with the importance (weighting) assigned to each of these factors. In this section, we will explore some of these top-level design drivers and consider their impact on the overall system design. These considerations are both technical and programmatic in nature; both aspects materially affect the viability of any system under consideration. While this listing is far from exhaustive, the parameters highlighted here tend to be common to every cellular satellite system design. Specifically, we will consider

- The User Terminal (UT)—The main interface the system user (customer) has to the system.
- The Target Market—Who do we expect the users to be, and how big is the potential user population?
- The Service Offerings—What are the users' expectations of the system, and features should we offer?
- The Operating Environment—The all-important link between the user and the satellite.
- The Value Proposition—If we build it, will they come? What business risks are we taking?

11.2.1 The User Terminal

The baseline assumption in this discussion is that the cellular satellite system users interface with the system using "disadvantaged" terminals. By this terminology we mean that the UT is "disadvantaged" from an electronic design standpoint, implying that the terminal is small, lightweight, and convenient to carry on the user's person. This convenience factor carries severe engineering penalties in the form of low antenna gain (impacts both uplink *and* downlink between the user and the satellite) and low transmit power (impacts the uplink). The system parameters most affected are the UT's gain-to-system-temperature (G/T) and its effective isotropic radiated power (EIRP); both the parameters are discussed further below and elsewhere in this handbook. The G/T must be high enough to ensure acceptable, low-noise reception of the signal from the satellite. Similarly, the EIRP must be high enough to ensure that the satellite will be able to reliably process and translate the signal from the UT. We refer to this basic design consideration as "closing the link."

UT convenience is far from being a "nice to have" characteristic in the commercial world of handheld wireless communications. This convenience plays a *significant* role in customer acceptance; a fact borne out within the Iridium cellular satellite system (discussed below). A small UT, in turn, makes the cellular

FIGURE 11.2 Military users accept bulkier UT form-factors.

satellite system engineer's job harder in that the lion's share of the system performance burden shifts to the space segment. The spacecraft, now, has to have both high EIRP and high-G/T in order that the subscriber can have a small, sleek terminal that easily fits in his pocket. Increased system complexity and cost is the result of this system trade.

There are niche applications where the UT trade result is more favorable to the system engineer. Specifically, military users are quite comfortable with bulkier UT form-factors (Figure 11.2) and are not averse to handsets that look and feel more like military-style transceivers. For the military market, we can employ a larger UT, with more antenna gain and more transmit power, thereby simplifying the spacecraft design task. The military user, in part, saved the Iridium system from being deorbited [1].

11.2.2 Target Market

Arguably, the most important parameter to identify is the market that the cellular satellite system intends to serve. Military needs tend to be very different from those typically considered for a commercial system. For instance, guaranteed availability, reliable communication through multiple layers of jungle canopy, nuclear event hardening, and low probability-of-intercept, which are often found in the requirements levied by the Department of Defense (DoD), are rarely, if ever, discussed in reference to commercial systems. Another consideration is whether the system is intended to serve a worldwide user group (including, perhaps, coverage of polar regions and open oceans), or the users are located in a general geographic region (e.g., Europe, Asia, specific landmasses). The state of terrestrial communication infrastructures in the target market—whether wired or wireless—must also be identified and evaluated as this could severely impact customer interest and service uptake.

These market considerations tend to drive several high-profile system design decisions. For instance, a DoD system, with a need to penetrate jungle canopy, would drive the selection of a particular operating frequency band to lower frequencies (e.g., UHF). Satellite orbit type would need to be of either the low earth or medium earth orbiting type (LEO or MEO) if worldwide coverage, including the polar regions, is required. Regional coverage, on the other hand, could be met with a satellite at geosynchronous earth orbit (GEO). The existence of adjacent terrestrial cellular coverage might influence the selection of a particular radio air interface structure (AIS). For example, if the satellite cellular coverage area includes population centers operating on a particular AIS (for instance the Global System for Mobile communication, or GSM), then it might be prudent to base the satellite AIS on something similar in order to simplify UT design and, in addition, facilitate roaming arrangements.

11.2.2.1 Expected Subscriber Population

An objective, well-researched estimate of the targeted subscriber population is an essential element to the viability of a cellular satellite system. Considerations include

- Expected total number of subscribers.
- Expected peak number of users within a given satellite footprint compared to the expected average user loading.
- Geographic distribution of the user population: whether they are evenly dispersed over the defined coverage area or concentrated in specific locations (population centers).

The total number of system subscribers drives ground-based design elements such as the capacity of the overall system authentication registers: the so-called home location registers (HLR) and the visitor location registers (VLR). On the other hand, the number of active users in the satellite field-of-view has a direct impact on the satellite architecture. Generally, the satellite downlink tends to be the limiting constraint on peak user handling. That is, the peak number of users that can be accommodated is based on the amount of available power in the downlink transmitter; since each downlink user signal receives a share of the total transmit power. In systems based on frequency division multiple access (FDMA), or a combination FDMA/TDMA (time division multiple access), it is important that the satellite transmit system is operated in the linear region of the transmitter gain characteristic. As more carriers are added to the composite (multicarrier) downlink signal (assuming equal power per carrier as necessary to close the link to the individual UT), the transmitter is driven closer to, and sometimes into, the nonlinear operating region. A transmitter, driven into its nonlinear operating point by a multicarrier signal, will generate intermodulation distortion (IMD) products due to the mixing of the individual carriers with one another (a classical heterodyning phenomenon).

In a system transmitting many (e.g., hundreds) of carriers, particularly when the carriers have modulation, the IMD tends to take on a random characteristic that manifests as a uniform noise spectrum that is band-limited by the output filter. The resulting effect is an increase in the transmitted system noise. This noise is characterized by the Noise Power Ratio (NPR) of the system and directly affects the downlink carrier-to-noise (C/N) ratio which, in turn, proportionately lowers the system capacity. System designs often include some type of automatic power control adjustment in order to maximize the number of downlink user signals that can be supported while maintaining the NPR at an acceptable level. Such a power control loop generally involves a measurement of received power level at the individual UT, which is then reported back to the Network Control Center (NCC) by way of a parallel control channel.

Assuming that the satellite transmitter can accommodate the expected peak traffic load, the next limitation becomes the switching capability within the satellite channelization equipment. The required switching capacity will generally dictate whether this channelization equipment is implemented in an analog or digital fashion. Lower capacity systems, such as the Inmarsat-3, are well served by conventional analog channelization and switching technology. In contrast, digital channelization and switching technology is required in modern cellular satellite systems that serve thousands of simultaneous users and have frequent associated call set-ups and teardowns. Although they do not support satellite cellular in the context considered here, the new (2006) Inmarsat-4 satellites, with their increased capacity over Inmarsat-3 (~16X per one Inmarsat publication [2]), include both digital channelizers and digital beamformers (DBF).

The distribution of users in the satellite field-of-view drives the design of the antenna (cell coverage) pattern. In satellite cellular systems, the bulk of which operate in the crowded L-band or S-band spectra, efficiency of spectrum use is a critical concern. Operators of a proposed cellular satellite system must apply for an allocation of the spectrum pool, usually as a result of common meetings of the World Radio Conference (WRC) wherein many operators negotiate for spectrum rights. Since, inevitably, granted allocations are less than that requested, a major motivation exists to design the proposed system with a high level of frequency reuse. Frequency reuse implies that multiple cells (beams) in the satellite system can reuse the same frequency but with different user signals modulated onto them. Typical reuse values, for regional systems such as ACeS and Thuraya, are on the order of 20 times. Thus, the same carrier frequency can be reused 20, or more, times over the satellite field-of-view thereby increasing the available capacity (the number of UTs that can be simultaneously served) by that same amount. By the same token, the link from the satellite to the Feeder, or Gateway, station must have an adequate spectrum allocation to accommodate the required terrestrial connectivity for the mobile UTs. Since Satellite-to-Feeder links are most often implemented in a higher frequency band (e.g., C-, Ku-, or Ka-band), wider operating bandwidths are generally more easily obtained and coordinated.

An operational scenario that includes many potentially active users located within a concentrated area (population center) can put a strain on the amount of available spectrum. This concentration can lead to the need for smaller cell sizes (narrower beam patterns) which, in turn, drives the size of the satellite antenna (making it larger and more difficult to accommodate on the spacecraft). Given a specific coverage area, or satellite field-of-view, smaller cells mean more cells and, therefore, more switching hardware on the spacecraft. This increase in hardware means that the spacecraft becomes more complex and costly.

If a system reaches its capacity limit, the user (the revenue source) experiences this through the busy signal. How frequently a user encounters a busy signal is referred to as "call blocking rate" and is one measure of the system's quality of service (QoS). The Quality of Service Forum [3] defines QoS, in part, as "consistent, predictable telecommunication service." Predictable service is what the subscriber demands and is, ultimately, the measure that will determine the success of the system.

The elements that enter into a system capacity assessment for a planned cellular satellite system include

- The number of call attempts per second
- The average call holding time
- The expected distribution of call types
  - Percentage of mobile originated (MO) calls
  - Percentage of mobile terminated (MT) calls
  - Percentage of mobile-to-mobile (MM) calls

All of these factors impact system design decisions regarding

- Satellite switching speed
- Satellite switching capacity
- Onboard buffering and memory capacity

11.2.3 Service Offerings

Another major system design driver relates to the types of services to be offered to the user community. Many of the cellular satellite systems currently fielded were designed, primarily, to provide voice communication services with some facility to provide low-rate data (9.6 kb/s or less) and facsimile services. The fact that some systems were not "future proofed" (i.e., physically incapable of supporting anything other than voice and low-rate data) has proven to be a considerable shortcoming and, in fact, has contributed to the business failure of at least one early system. The explosive advent of the Internet in recent years has given rise to major advances in the wired communications infrastructure on a worldwide basis. The ubiquity of data communications at virtually every social stratum has fueled the tremendous growth of so-called e-commerce, a market that, in 2005, was valued at $8.5 trillion [4]; vastly exceeding all expectations [5]. Of this huge market, a very significant \$121 billion slice is predicted (for 2008) to be transacted over wireless handsets alone [6]. Consequently, there exists a large incentive for a satellite cellular system operator to ensure that the system is capable of supporting wireless data services at attractive data rates (on the order of 100 kb/s or more).

If the system is to support data services, then there is a decision to be made as to whether the services will be sold as circuit switched or packet switched. Circuit switched services represent the traditional approach to providing communication services: a channel is allocated, and dedicated, to the customer for the duration of the call. The customer is subsequently billed for the amount of time that the call was active, whether or not any data were passed during that time. Packet switching, on the other hand, is metered on the basis of the amount of data that are transmitted. The customer does not use any system resources while idle. It is a "service on-demand" paradigm in a mode of being "always connected" (similar to having a computer on a local area network in the wired world). Compatibility with data transmission frameworks, such as the transmission control protocol/internet protocol (TCP/IP), is more streamlined and more efficient in a packet-based system. Infrastructures such as those developed for the general packet radio service (GPRS) need to be incorporated into the design of the cellular satellite system if it is to support packet switching.

In summary, decisions on service offerings involve considerations as to whether the service backbone is to be voice-only or voice and data. If data are included, then the choice between circuit switching and packet switching needs to be considered. In addition to the network equipment complement needed to execute these services, these decisions drive the basic system parameters of

- *Air interface*: TDMA, FDMA, CDMA.
- *Channel spacing*: Higher data rates require wider spectral bandwidths which, in turn, mean that the channels need to be spaced further apart. CDMA approaches, likewise, require wider spread bandwidths for higher data rates.
- *Modulation type*: Bandwidth-efficient modulation methods should be employed in order to conserve system resources by transmitting the most amount of data for the lowest power and the narrowest bandwidth. Multilevel modulation types can be traded, such as various types of phase shift key (PSK) or quadrature amplitude modulation (QAM).

11.2.4 Operating Environment

For a commercial system, putting aside political issues such as frequency coordination for the moment, we would want to select an operating frequency band that can penetrate at least *some* buildings. In addition, the operating band should support acceptable performance with nondirectional antennas on the UT (i.e., provide tolerably low path loss with reasonably sized electrically small antennas). These constraints tend to drive toward lower frequency bands (e.g., L-band or S-band, which are roughly 1500 MHz and 2500 MHz, respectively).

11.2.4.1 Link Margin Considerations

During the system design phase, the radio link between the satellite and the UT has to be given a great deal of detailed consideration. Both the typical and disadvantaged user conditions need to be handled. For instance, the typical case might find the user with a clear line-of-sight to the satellite. On the other hand, a disadvantaged condition might be where the user is in the middle of an office building, in a city (a "concrete jungle" where multipath propagation could be an issue) and, perhaps, at the edge of the satellite's coverage area (low elevation angle, greatest path distance to the satellite and, therefore, greatest path loss). Further, meteorological conditions such as rain and, to a lesser extent, snow will degrade the link performance in proportion to both the rate of precipitation and the frequency of operation. Operation in a tropical area would clearly be more affected by rain attenuation than if the system were intended to serve, say, Northern Africa. In all of this, link margin is "king." If we take the traditional approach to assessing the amount of available link margin in the system, we can define link margin (nonrigorously, and in decibel terms) as

$$\text{Link Margin} = (C/N)_{\text{received}} - (C/N)_{\text{required}},$$

where

- "$(C/N)_{\text{received}}$" is the ratio of the power of the transmitted signal (at the receiver) to the total noise power (e.g., thermal noise, interference, and other degradations) impinging on the receiver. In the example of the link going to the UT, the satellite is the transmitter and the UT is the receiver. Mobile service providers will often refer to this transmission direction as the "Forward" direction. Transmission from the UT to the satellite is, likewise, called the "Return" transmission.
- "$(C/N)_{\text{required}}$" is the minimum ratio of received signal to total noise required at the receiver in order to accomplish acceptable detection with the modulation method selected.

We can gain additional insight into what this equation means by breaking $(C/N)_{\text{received}}$ into its constituent components. Thus

$$\text{Link Margin} = \{\text{EIRP} - L_{\text{impairments}} + (G/T)\} - (C/N)_{\text{required}},$$

where

- "EIRP" is the Effective Isotropic Radiated Power of the transmitter (e.g., the satellite in the case of the link from the satellite to the UT). In "rule-of-thumb" terms, EIRP consists of the gain of the transmit antenna multiplied by the RF power applied to its input port.
- "(G/T)" is the figure-of-merit often applied to satellite receiving stations, in this case the UT. (G/T) is the ratio of the passive antenna gain (G), at the receiving station in the direction of the incoming signal, divided by the total system noise of the station expressed as an equivalent temperature (T). The primary component of the system noise is, usually, the noise added by the passive components (following the antenna) and the noise figure of the first amplifier, or low-noise amplifier (LNA), in the receiver. In addition, the noise contributed by the subsequent components in the UT receiving chain is referred back to the antenna terminal (suitably scaled by the gains of the components ahead in the signal flow).
- "$L_{\text{impairments}}$" is the combination of losses and effects on the transmission channel that tend to degrade the overall received signal quality.

In other words, link margin can be thought of as the excess desired signal power available at the receiver once we have accounted for the signal strength required by the demodulator in the receiver (commensurate with the chosen modulation format), the inevitable thermal noise, and the impairments suffered along the way. Some of the impairments that have the greatest detrimental impact on the link margin are

- Transmitter intermodulation, as characterized by the NPR.
- *Spreading loss*: path loss, which is proportional to the distance between the UT and the satellite; very much driven by the slant range, or elevation angle at the UT, and the transmission frequency.
- *Atmospheric loss*: signal absorption in the propagation path proportional to the carrier frequency and the amount of moisture in the air.
- *Polarization mismatch*: orientation of the UT antenna relative to the satellite antenna.
- *Body losses*: absorption and blockage of the communication signal by the user's body.
- *Multipath interference*: in CDMA systems, rake receivers can actually take advantage of multipath to *enhance* the received signal. In most other multiple access techniques, multipath is detrimental.
- *Cochannel interference*: leakage from the other channels on the same frequency but in a different beam (or, for a CDMA-based system, signals in the same beam but with a different code).
- *Adjacent channel interference*: leakage from a channel on an adjacent frequency.

- *Digital implementation loss*: effects such as spectrum truncation due to the finite filtering bandwidth in the receiver's demodulator.

Each of these factors must be quantified and accounted for in the overall impairment budget in order to ensure that the link margin needed to provide acceptable service is attained.

11.2.4.2 Implementation Loss

It is of particular interest to consider link margin in the light of the modern trend toward digital satellite systems where a substantial amount of onboard signal processing takes place. Many of these digitally processed systems will demultiplex, demodulate, process, switch, remodulate, and remultiplex the communication signals. These digital transponders stand in contrast to traditional "bent pipe" transponder approaches where the uplinked signal is merely filtered and translated in frequency before being downlinked. In systems that are all-digital, the primary performance measure of interest is the demodulated E_b/N_0, which is the signal energy per bit (E_b) divided by the noise density (N_0). We can relate the classical signal (or carrier) to noise ratio to E_b/N_0 by way of the raw transmission bit rate and the pre-detection bandwidth, as

$$(C/N)_{dB} = (E_b/N_0)_{dB} + (R/W)_{dB},$$

where R is the bit rate in bits per second and W is the pre-detection bandwidth in Hz.

Typical performance curves plot bit error rate as a function of E_b/N_0, the so-called waterfall curves, as shown in Figure 11.3.

Each impairment that we discussed above could be considered as a contributor to the overall digital implementation loss. In effect, for digital modulation schemes [e.g., m-ary PSK or n-ary QAM, where m and n are integers] there are three contributors to implementation loss that can be quantified and modeled as a characteristic of every block in the overall system. These contributors are

- *Additive white Gaussian noise* (*AWGN*): Thermal noise, random interference, and other random noise effects that degrade the received signal-to-noise ratio (and, therefore, directly affect the E_b/N_0
- *Phase distortion*: Nonlinearities in the transmission of phase information
- *Amplitude distortion*: Nonlinearities in the transmission of amplitude information

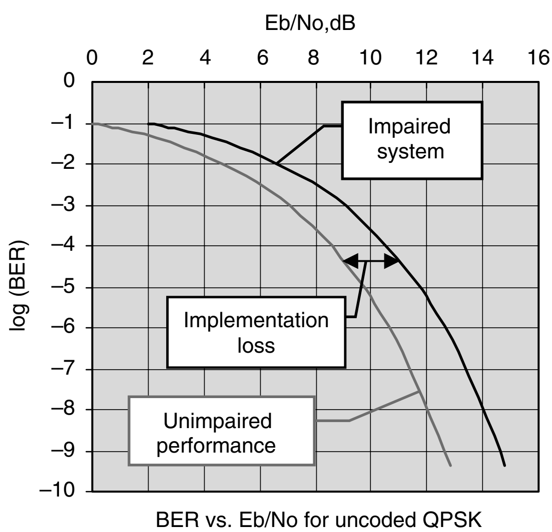


FIGURE 11.3 Typical "waterfall curve"; bit error rate versus Eb/No for uncoded QPSK.

In fact, all of the impairments we have discussed could be couched within the above three terms. The amount of implementation loss suffered, at any given stage, will be a strong function of the type of modulation chosen. For instance, *m*-ary PSK has the intelligence coded in the phase of the signal only and will be much more affected by phase distortion than by amplitude distortion. Generally speaking, one can hard-limit a PSK signal and not lose its detectability. In contrast, an *n*-ary QAM signal uses both amplitude and phase to code the intelligence. Such a signal would, clearly, be sensitive to both amplitude distortion as well as phase distortion. Each contributor to the overall implementation loss tends to shift the "waterfall curve" to the right (as shown in Figure 11.3), implying a need for higher E_b/N_0 for a given (desired) bit error rate.

Consequently, for digital systems there is justification for assessing link margin solely in terms of the various implementation losses encountered throughout the system. That is not to say that the conventional measures we previously discussed are obsolete. On the contrary, if we have limited control over the signals to be carried within the cellular satellite system we are designing, then the conventional measures offer the best common ground on which to specify the system performance. On the other hand, if we have complete control over both the ground and the space segments of the system, and it is an all-digital system (with a modulation format of our own choosing), we can streamline the entire analysis process, and drive more directly to the bottom line (i.e., E_b/N_0), if we model each segment in terms of its digital implementation loss impact on the end-to-end performance [7].

The success of a given system hinges on customer satisfaction, and a large part of that satisfaction is derived from the ability to make successful calls "most of the time." The larger the link margin (i.e., the more degradation the link can suffer before communication is no longer possible) the more frequently a user will be able to complete successful calls, and the more often he is likely to use the system. The link design has to contain enough margin to cope with these "most of the time" situations, which is why cellular satellite system links are, as a rule, designed on a statistical basis [8]. No hard and fast rule exists here as the trade-off is subject to interrelated and, often, subjective criteria. Current cellular satellite systems, regardless of whether LEO, MEO, or GEO, tend to present the "average" user with about 10 dB of link margin after all impairments have been considered. This, somewhat counterintuitive result (given that GEO constellation orbits are some 40–50 times higher than the orbits of LEO systems) is due to the fact that the lower-complexity LEO satellites, as compared to GEO satellites, are smaller (lower overall transmit power capability) and must cover a broader angular field-of-view (lower antenna gain).

If sound engineering conservatism prevails against the, sometimes exuberant, optimism of the marketing department, the "average" user will be defined as one who is located closer to the edge of a typical cell beam (rather than at the peak), does not have his UT antenna ideally oriented with respect to the spacecraft, and will probably be located inside of a building (but not too far away from a window). Such a design approach will tend to meet the "most of the time" criterion.

We can apply other methods to address 5–10% of the conditions when the user is not ideally positioned. These methods generally include the assumption of a cooperative user. As an example, for a mobile terminated (MT) call (i.e., one in which the mobile user has an incoming call) the paging signal can be issued at a higher signal strength which, in turn, requests the user to move to a better line-of-sight position with respect to the satellite.

11.2.4.3 Latency

Latency refers to the amount of communication delay a cellular satellite system user experiences during a conversation or a data transaction. The effect of "substantial" latency can range from mild irritation (users at each end of the link talking over one another) to major transmission inefficiencies (multiple retransmissions under an IP environment), including dropped transactions, which could be disastrous at several levels ranging from economic to human safety considerations. The latency equation contains several contributing factors including

- *Propagation delay*: the transit time (determined by the speed of light) for the signal to travel between the satellite and the UT. The total delay is, necessarily, twice the one-way delay since the signal must

go from earth to satellite and back down again. The approximate round trip (earth–satellite–earth) delay for the three main orbital configurations are
  - GEO (orbital altitude of approximately 35,900 km): 239 ms.
  - MEO (orbital altitude of approximately 10,300 km): 69 ms.
  - LEO (orbital altitude of approximately 700–1400 km): 6 ms.
- *Coding delay*: in digital systems the data to be communicated (be it digitized voice or data files) is coded for various reasons (error correction and data compression are the most common reasons beside encryption and security concerns). The coding process implies that blocks of data must be stored before coding. Received blocks must, likewise, be buffered in blocks before decoding can take place. This buffering/coding process adds delay to the communication transmission. The amount of coding delay added to the transmission varies according to the number of coding layers and the type of voice coder–decoder (codec or vocoder) employed. This coding delay can easily range between 50 and 150 ms.
- *Relay delays*: There may be a need for a particular call to be routed through multiple nodes [for instance, satellite-to-satellite intersatellite links (ISLs), ground station-to-ground station, or relay station-to-relay station]. Sometimes channels are demultiplexed and remultiplexed at these intermediate points (depending on the routing required and the multiplexing hierarchy required to effect the relay path). Again, the magnitude of the delay depends on the path taken, but it will generally range between 10 and 40 ms.
- *Other system delays*: Depending on the system considered, other delays could also be applicable. Some of these are listed below.
  - Aggregated digital communication channels, in the terrestrial infrastructure, are frequently buffered for purposes such as synchronization. In addition, echo cancellers and the relative placement of the channel carrier (e.g., especially if placed close to the band edge of the channel filters) contribute varying amounts of delay. These processes also contribute to the overall latency that the user experiences and can range from negligible to 120 ms (for the worst combinations of echo cancellers and channel filter group delay).
  - Emerging terrestrial systems are being designed to accommodate network protocol infrastructures such as H.323 or SIP, which include support for "Voice over IP" (VoIP). Considerations, such as variable grades of service, are integral to these protocols and imply additional latency contributions. These considerations will apply directly to cellular satellite systems that are configured to operate in compatible packetized modes.
  - Concerns about the impact of latency on Asynchronous Transmission Mode (ATM) and TCP/IP communications over the satellite channel have interested many workers in the field. Many different approaches have been analyzed and experimentally evaluated to deal with the transmission inefficiencies that, potentially, could arise [9–12]. Current and future cellular satellite systems will need to have the capability of accommodating bursty packet modes such as these.

The relative impact of the various latency factors will, clearly, vary with the type of system. For a GEO system propagation delay generally dominates the latency characteristic, resulting from the higher altitude of the satellite. LEO and MEO systems, with their lower orbital altitudes, have corresponding lower values of propagation delay. This factor is one of the main reasons that several cellular satellite system designers have chosen LEO or MEO constellations. On the basis of propagation delay alone, the decision would appear to be a "no-brainer" in favor of MEO or, especially LEO (particularly since the lower altitudes also provide a potential for lower spreading loss and, possibly, greater link margin). The subtleties of the other latency contributors, however, could easily conspire to erase the apparent advantage of the lower altitudes. For example, consider two users with LEO cellular satellite service located at opposite extremes of a nominal GEO coverage area (separated by, say, 4000 km). If we assume that the two systems (LEO and GEO) have equal coding delays, the LEO path might have to traverse as many as four satellite planes or, perhaps, an equal number of intermediate ground stations (if the system does not include ISLs). Consequently, with each relay making its own contribution to the overall latency, it is easy to see how a point-to-point LEO communication could enter the realm of the GEO propagation delay. This discussion is not intended to favor one type of system over another with respect to latency, but merely to point out that one must consider the full complement of effects in carrying out an impartial trade study.

11.2.4.4 Orbit Altitude and the Van Allen Radiation Belts

Though considerations of latency, individual satellite coverage and link margin all tend to enter into the selection of system orbit altitude, another physical constraint exists that impacts the placement of the specific orbit. This constraint is due to the location of the Van Allen radiation belts. These radiation belts (first discovered and characterized by Dr. James Van Allen in 1959) consist of two annular rings of radiation that encircle the earth and are centered on a plane defined by the equatorial latitude [13,14]. Sensitive spacecraft electronic equipment is readily damaged by high doses of radioactivity. Mitigation techniques usually involve component shielding of one form or another (e.g., thick metal boxes augmented with spot shielding composed of strips of tantalum), all of which imply additional spacecraft mass. Since mass is a premium item in spacecraft design (launch costs are proportional to launch mass), it is, clearly, desirable to make the satellite as light as possible. Flying a satellite within the Van Allen belts would entail massive amounts of shielding in order to achieve reasonable mission life. Consequently, orbits tend to be placed either to avoid the Van Allen belts altogether or to make only occasional (and very rapid) transitions through them. A schematic view of the radiation belt locations, from a polar perspective, is shown in Figure 11.4. This figure also shows the relative locations of the three most common satellite orbits.

Other orbit types have been proposed for use in cellular satellite designs. For instance, in the late 1990s there was a proposed satellite cellular system called Ellipso [15]. In this system, a pair of elliptical orbits (perigee of 520 km and apogee of 7800 km) was supposed to slip between the two Van Allen belts during the north-south transit.

11.2.4.5 Operating Environment Summary

In evaluating the operating environment of the system, we are primarily interested in assessing the amount of available link margin and in determining the best way to maximize it. Higher link margin primarily impacts

- The size of the satellite antenna (dramatically impacting cost, mass, and complexity)
- The optimum frequency band of operation (which, more often than not, has political ramifications that swamp the technical considerations)

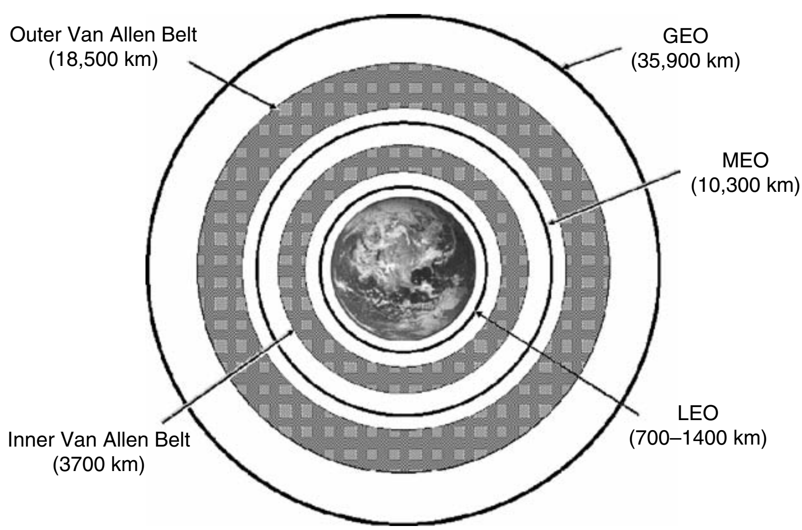


FIGURE 11.4 Locations of Van Allen radiation belts and typical cellular satellite orbits.

- The power and linearity performance of the satellite transmitter (also, a major cost element of the system)
- The noise performance (sensitivity) of the satellite LNA and receive system

Where system latency is determined to be of high priority with respect to cellular satellite system performance, the main impacted design parameters tend to be

- *Satellite orbit altitude*: impacts propagation delay, which tends to exhibit the greatest variability of all of the components of the latency equation
- *System coding approach*: while important with respect to the latency assessment, may be more strongly driven by requirements to provide error detection and correction (EDAC) as well as communication security

11.2.5 Value Proposition

Will the typical user, of the cellular satellite system, be a business traveler or will the system business case assume a broader user population (penetrating lower economic strata)? Experience has shown that a target subscriber population based, primarily, on the affluent, worldwide business traveler places costly constraints on the system design and may not, necessarily, be attractive to the targeted consumer. For one thing, such a business case requires a worldwide coverage of satellites. This kind of coverage, in turn, dictates many satellites with narrow beams covering the entire surface of the earth. The most obvious approach to such worldwide coverage is to deploy a large fleet of satellites in a LEO configuration. The LEO constellation, if properly configured, provides the desired continuous global coverage, but it turns out to be extremely expensive to deploy such a system. Although the individual satellites, in a LEO system, tend to be small and inexpensive (in a relative sense considering the overall context of spacecraft technology), many satellites are required in order to achieve the desired coverage (tens to hundreds). The network infrastructure also demands a large number of gateways, with active features to track the rapidly moving satellites, in order to handle the connection of the mobile traffic to the public wired backbones. Clearly, the satellites could be designed to perform the call-by-call traffic switching; including relaying between satellites via ISLs, which, in turn, reduces the quantity of required gateways, but this approach adds complexity to the space segment and increases the overall system cost. In addition, it is almost a necessity to have the entire constellation in place before a reasonable level of service can be offered for sale. A gradual ramp-up in service, on the basis of a partially deployed constellation, is very difficult (if not impossible).

A system design based on the MEO configuration also has the potential of providing worldwide, or nearly worldwide, coverage with fewer satellites (on the order of tens or less) relative to a LEO version. MEO spacecraft tend to be more complex than LEO spacecraft, but less complex than their GEO counterparts. Fewer gateway stations are required to service a MEO constellation and, since the spacecraft are at a higher altitude, they move relatively slowly so that tracking is easier. For the same reason, call hand-offs between spacecraft tend to be less frequent. As a consequence, the total system cost for a MEO system tends to be somewhat lower than the cost of an equivalent LEO system. In addition, depending on the actual orbits designed into the system and the quantity of spacecraft per orbital plane, MEO systems have the possibility of offering start-up service to selected areas on the earth. Thus, some revenue can be returned to the enterprise prior to the complete deployment of the satellite constellation.

Total system cost is a major issue with regard to service pricing. The price point is generally set so as to provide a profitable return within a set timeframe consistent with the business plan of the enterprise. It is evident that the base system cost is a large part of the initial investment against which a threshold to profitability is set. Obviously, the sooner the system finances cross this profitability threshold the sooner the system can be declared to be a success, making for a happy investor community. In this regard, GEO systems tend to be more cost-effective, with a greater probability of being profitable at a lower price point. Although GEO spacecraft tend to be larger, more complex, and more costly than those destined for either LEO or MEO service, only one gateway is required to service the communication links. Connectivity is instantly available with the arrival of the spacecraft on station, and commercial service can generally be offered within a few months thereafter. Although GEO systems are, by nature, regional, careful selection of both service region and market mean that revenues can start flowing relatively quickly. Since the overall cost of deploying a regional GEO cellular satellite system is, by and large, lower than either a LEO or a MEO system, the threshold to profitability is potentially closer. The downside is, of course, that regional GEO systems, individually, cannot provide global coverage. Theoretically (except for the Polar Regions), three spacecraft placed at 120° longitude intervals can provide global coverage. The requisite narrow beams needed to cover the earth field-of-view would, however, require multiple, very large, and possibly impractical, antennas in order to close the link to UTs. Alternatively, one could take the approach of concentrating on the major potential revenue producing areas on the earth's surface (major landmasses), largely ignoring the open ocean areas. Under such a scenario "earth coverage" could be accomplished with four to six GEO spacecraft, with two collocated spacecraft at longitudes over North–South America, and over Scandinavia–Europe–Africa.

Regardless of the approach taken, a successful cellular satellite system value proposition hinges on meeting the defined technical and service requirements with a system embodiment that minimizes system implementation cost and maximizes return on investment. The probability of obtaining a positive return on investment is inversely proportional to the risk taken in the design and deployment of the system. Several approaches can be taken to reduce the overall risk of the project. These risk reduction approaches include [16]

- *Minimizing development cost*: reuse previously qualified designs to the maximum extent possible without compromising required system performance. Maximum reuse of qualified hardware also enhances (shortens) project schedules, and speeds system deployment and time-to-market.
- *Maximizing system flexibility*: anticipate future developments and plan to accommodate them. For instance, future data services will require wider channel bandwidths. A design that includes wider channel bandwidths will serve both current needs (e.g., voice transmission) in addition to allowing system migration to data services in the future.
- *Well-researched business plan*: diligently considered markets, along with a firm financing plan, helps to ensure steady progress in the deployment of the system.

11.3 Approaches

Having considered the fundamental driving factors of a cellular satellite system design, we now turn to specific implementations and the trade-off considerations that they imply. For convenience, a summary of the driving parameters considered up till now is shown in Table 11.1.

Regardless of the system embodiment selected, certain elements are common and will be found in any cellular satellite system. These main elements are depicted in Figure 11.5.

Technology will, inevitably, continue to evolve. Unfortunately, once a satellite system is launched it is generally impractical to make modifications to the space segment. The typical 10–15 year on-orbit design lifetime of a spacecraft is close to an eternity when viewed against the backdrop of historic technology trends. These two incompatible facts make it incumbent upon system designers to anticipate the future and strive to make accommodations for upgrades to the extent practical. As we have previously discussed, system flexibility is key to future proofing the design. Flexibility allows the system to support evolving services with minimum modification. For instance, a space segment able to support wideband channels will be able to grow with the relentless trend toward higher data rates. Beamforming flexibility (allowing the coverage area to evolve with the market) is another approach to future proofing, either by way of digitally programmed onboard beamforming, ground-based beamforming, or by way of an overdesigned conventional analog approach. In the next section, we will look at some specific system designs, the embodiments of which are the results of decisions made by previous cellular satellite system designers after having grappled with the concepts we have discussed.

TABLE 11.1 Summary of Cellular Satellite System Driving Parameters

Driving Parameter	Trade Issue
User terminal convenience	Satellite EIRP and G/T
Target market and its location	Operating frequency
	Orbit type
	Air interface structure
Subscriber population and geographic distribution	System capacity
	Authentication register size
	Satellite aggregate EIRP
	Multiple access method
	Satellite linearity requirements
	Satellite switching speed
	Satellite switching capacity
	Satellite memory capacity
	Channelization approach
	Degree of frequency reuse
	Cell size
Operating environment	Multiple access method
	Satellite EIRP and G/T
	Modulation method
	Phase and amplitude linearity
	Available link margin
	Orbit type
	Latency
	Band of operation
	Coding method (digital system)
	High penetration alerting method
Service offerings	Air interface structure
	Channel spacing
	Channel bandwidth
	Modulation type
Value proposition (profitability)	Overall system cost
	Risk element (amount of new technology)
	Incremental revenue possibilities
	System flexibility (future proofing)
	Business plan
	Funding base

11.4 Example Architectures

11.4.1 LEO

In this section, we will examine two of the so-called Big LEO cellular satellite systems. These systems are "Big" since they have relatively large complex spacecraft and are designed to handle large quantities of information. These characteristics stand in contrast to the antithetical "Little LEOs" whose main mission is to provide short message service, paging, and asset tracking, and are not considered cellular satellite systems for our purposes here. Common to all LEO cellular satellite systems is a large quantity of satellites in orbits that range from 700 km to about 1400 km (avoiding the inner Van Allen belt). LEO satellites also tend to be the simplest (or, at least, the lightest) of the major communication satellite types, and tend to have the shortest service life (5–7 years) due to limited onboard capacity for station keeping fuel. Although somewhat counterintuitive, Big LEO systems tend to be the most costly to deploy (of the LEO, MEO, GEO varieties), mainly due to the large quantity of satellites required and the extensive globally distributed ground infrastructure required to support the system. The main marketing point for LEOs is

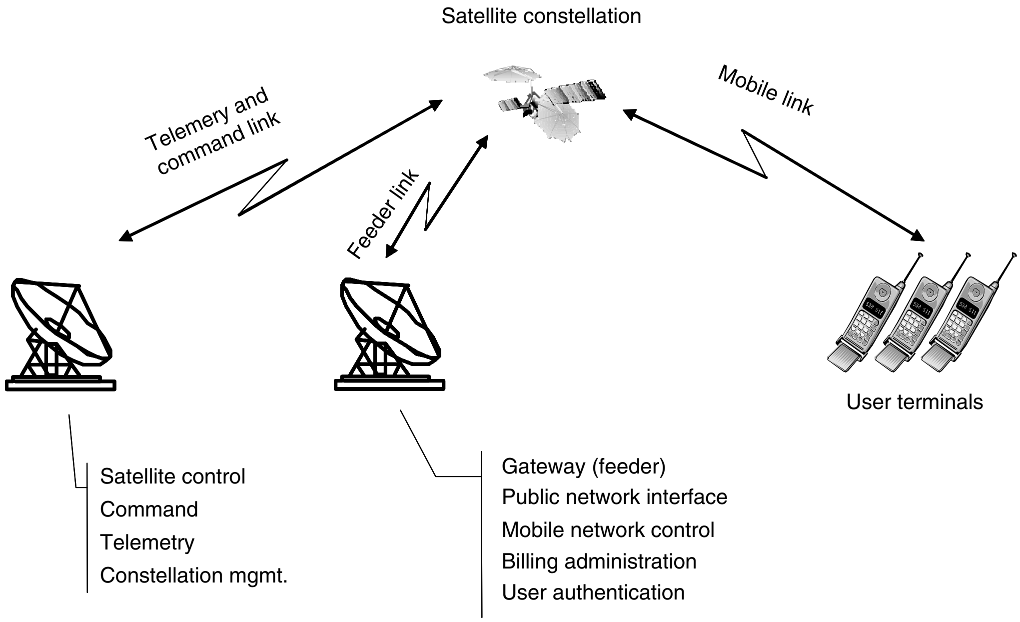


FIGURE 11.5 Common elements of a cellular satellite system.

the low latency between the ground and the satellite due to the close proximity of the satellite to the user. The two best known of the Big LEOs are Iridium and Globalstar.

11.4.1.1 Iridium [17,18]

When launched, the Iridium cellular satellite system was owned by Iridium, LLC, of which Motorola was an 18% owner. The main contractors who supplied the system hardware were Raytheon (main mission antennas), Lockheed Martin (spacecraft bus), and Scientific Atlanta (Earth terminal equipment). Sixty-six active satellites, in six polar orbital planes of eleven satellites each, ring the earth at an altitude of 780 km. The Iridium system was originally targeted at providing, primarily, voice service to the globetrotting business professional (more on this later). However, the system is also designed to support low-rate data communications as well as facsimile and paging. The system is now owned by Iridium Satellite LLC, the company that purchased the bankrupt system in November, 2000 [19]. The Boeing company operates the system for Iridium Satellite LLC.

Iridium uses a protocol stack that is partially built on GSM and partially Iridium unique. Therefore, the system is compatible with the GSM infrastructure at the service level, even though the physical layer (the radio link) is not in accordance with GSM standards. The main elements of the Iridium system are shown in Figure 11.6.

The main mission links (to the UTs) is accomplished by a combined FDMA/TDMA time division duplex (TDD) method in the L-band (specifically, 1610–1626.5 MHz) and with QPSK modulation. Voice communication is digitized in a vocoder, and the individual voice channels operate at a data rate of 2.4 kb/s. Traffic is passed between the individual satellites on cross-links for the purposes of traffic routing and call hand-off, as one satellite transits out of view and another is needed to pick up the connection. These cross-links are operated at K-band (23 GHz), but at a higher data rate of 25 Mb/s. One of the motivations for including the complexities of cross-links in the spacecraft design was to enable efficient call routing (less dependent on the physical location of the gateways) which, in turn, enabled full service coverage to the oceanic regions.

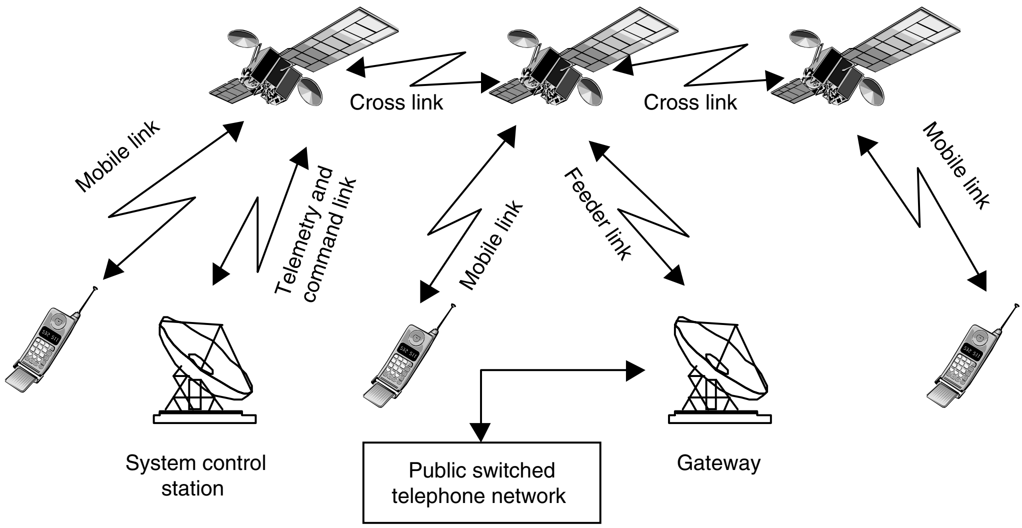


FIGURE 11.6 Elements of the Iridium communication system.

Globally distributed gateways, each with a 3.3 m antenna, provide the interface between the Iridium system and the public wired networks. Links between the satellites and these stations is carried out via the spacecraft feeder antennas on a K-band carrier (19 GHz down and 28 GHz up on QPSK modulation at a coded data rate of 6.25 Mb/s). Call setup and teardown is controlled at these local gateways, and this is also where billing records are generated. Each gateway also includes the necessary user authentication equipment (HLR, VLR) as well as the necessary infrastructure to enable UT position location. Position location is important for political reasons, among other things, so that local jurisdictions can maintain control over telecommunication traffic in their respective regions. At least two gateway antennas are required at each site in order to properly track and smoothly maintain contact with spacecraft in the field-of-view.

Management of the whole system is carried out at the System Control Station, which is physically located in Lansdowne, Virginia. Here the network infrastructure is monitored and controlled. This station also takes care of satellite maintenance and control, monitors status and system health, and serves as the center for any troubleshooting required.

The 700-kg Iridium spacecraft is very sophisticated. Its design includes a complex digital signal processor that demodulates incoming signals, switches them and remodulates them as needed. One great advantage of this method is that uplink implementation loss can be significantly isolated from the downlink, thereby improving the overall bit error rate performance. The downside to this sophistication is, naturally, a major increase in the complexity and cost of the spacecraft. Three main mission antennas form the 48 cellular beams per satellite by way of direct radiating phased array technology. These beams cover a footprint on the earth some 4700-km across. Each satellite has the switching capacity to handle 3840 simultaneous calls, but power considerations limit the practical number to around 1100.

As alluded in the opening paragraph of this section, the Iridium system fell on hard times economically. Service was discontinued for a time, and plans were being made to actually deorbit the satellite constellation. This system is an object lesson in how even the most sophisticated, well-engineered system [20] can become a failure if mishandled from a business perspective. Of the many problems that beset the Iridium system, some of the more significant ones include

- *Market share erosion*: at the time of conception (1988–1990) the enormous build-out of inexpensive terrestrial cellular was not anticipated. This problem is the same one that contributed to the lackluster business performance of the American Mobile Satellite system.

- *High system cost*: estimates vary, but the cost of the Iridium system was some $5–$7 billion. This high cost, in turn, implied the need for high service cost ($3–$5 per minute) if returns were to be realized according to the business plan schedule. In addition, the acquisition cost of the UT to the subscriber was high ($2000–$3000). This combination of high service and terminal costs put the system out of reach to all but the most affluent consumers. At this writing, the new Iridium owners have a number of partner service providers offering UTs for around $1500 with airtime rates on the order of $1.20.
- *Market overestimation*: Iridium targeted the global business traveler who would often be outside of terrestrial cellular coverage. Most high-intensity business destinations today are well served by inexpensive terrestrial cellular coverage, thereby diminishing the need for Iridium's ubiquitous coverage.
- *Ineffective marketing*: Obtaining Iridium service was not a streamlined process. In addition, advertising was underwhelming and tended to depict users in polar ice fields or deserts (a very limited revenue population).
- *Poor UT form factor*: The Iridium UT has been variously described as a "brick" or a "club." The bottom line is that the UT was awkward and heavy in a user environment used to small pocket-sized cellular telephones. The UT was designed by engineers with very limited regard to user appeal and aesthetics.

Although the original investors in the Iridium system did not see the anticipated return on their investments, the system continues to operate as of this writing (2006). The new owner purportedly acquired the system for "pennies on the dollar." Therefore, with a lower acquisition cost advantage, the business model now closes. The heaviest user is the US DoD who has placed a dedicated gateway in Hawaii. Iridium's design allows the DoD to have global, secure communications coverage with full control over the traffic. No communication is downlinked to the ground anywhere except the Hawaii gateway, thanks to the satellite-to-satellite cross-link design. However, in general, Iridium is focusing on niche markets that are not covered by ubiquitous cellular or landline coverage, as evidenced by a statement posted on their website in 2006:

> If you work, live, or travel in areas outside cellular coverage or in areas with inadequate landline service, Iridium provides an immediate solution. Eighty-six percent of the world's landmass and all of its oceans are in areas with inadequate landline service. Iridium addresses this situation by providing coverage in all ocean areas, air routes, and all landmasses—even the Poles.

11.4.1.2 Globalstar [19,21]

Globalstar is another Big LEO system that has been fielded. At its inception, this system was owned by a partnership of several well-known companies led by Loral and Qualcomm as the general partners. System hardware suppliers included Alcatel (gateway equipment) and Alenia (satellite integration). The company filed for bankruptcy in February, 2002, and emerged from bankruptcy in April, 2004. The acquiring company was Thermo Capital Partners LLC, who reportedly purchased Globalstar for $43 million [22].

The Globalstar system consists of 48 satellites, in eight planes of six active satellites apiece, flying at an altitude of 1410 km. Thus, the Globalstar orbits are the highest of the Big LEOs fielded to date, which also means that the individual satellites move more slowly and are easier to track at the gateway stations. Each orbit is inclined at 52°, thereby concentrating service resources in a band bordered by the 70° north and south latitudes. These latitude limits were carefully chosen in that the majority of the Earth's population is located in this region (the "revenue" latitudes). The system uses a CDMA air interface that is based on the popular IS-95 terrestrial standard, which includes such features as soft hand-offs, dynamic power control (essential to preserve capacity in a CDMA system, and preserves handset battery life in the bargain), and soft capacity limits (2000–3000 simultaneous calls per satellite). The intelligence signal is spread to 1.25 MHz, and these CDMA channels are spaced, in an FDMA fashion, on 1.23 MHz centers. Globalstar targeted, pretty much, the same customer base as Iridium went after (i.e., the global business traveler),

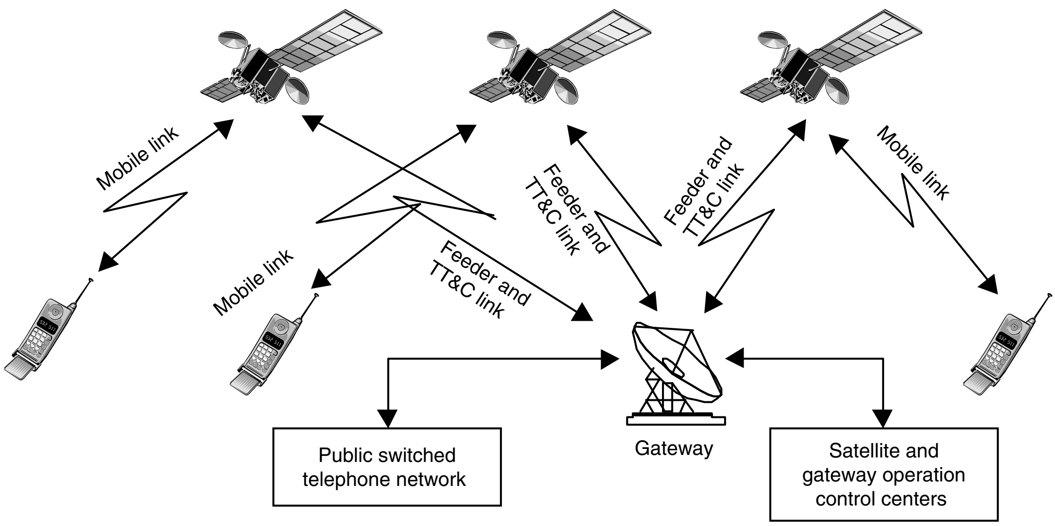


FIGURE 11.7 Elements of the Globalstar communication system.

and offers an equivalent suite of services. For instance, the Globalstar offerings include variable-rate voice (2.4, 4.8, and 9.6 kb/s), low-rate data (on the order of 7.2 kb/s), facsimile, paging, and position location.

Mobile links are realized at S-band down (2483.5–2500.0 MHz) and L-band up (1610.0–1626.5 MHz) with QPSK modulation. The satellites are simple, bent-pipe transponders (amplification and translation only) with gateway links (satellite—gateway) at C-band (5019–5250 MHz up, and 6875–7055 MHz down). Rake receivers are included in both the UTs and the gateways in order to use multipath signals to advantage in strengthening the links. As a result, typical Forward E_b/N_0 (the weakest link due to satellite power limitations) is reported to be on the order of 4 dB, while the return link weighs in at around 6 dB.

The main elements of the Globalstar system are depicted in Figure 11.7.

The Gateway Operations Control Center (GOCC) and the Satellite Operations Control Center (SOCC) are collocated in San Jose, CA. The individual, globally distributed gateways, with their 5.5-m tracking antennas, perform the local functions of the network administration. This administration includes user authentication (HLR, VLR functions), call control and local switching, local PSTN connections, and satellite TT&C processing.

Each satellite has a pair of direct radiating, fixed-beam phased array antennas to form the L-band cellular beam patterns. The transmit antenna consists of 91 elements, each having its own SSPA, phased with a fixed beamformer into 16 transmit beams. The 16 receive beams (which are congruent with the transmit beams) are, similarly formed with a fixed, low-power beamforming system, but the array consists of only 61 elements, each with its own LNA. The graceful degradation properties inherent in a phased array mean that redundant SSPAs and LNAs are not required. The satellites have a design lifetime of between 7 and 8 years. Figure 11.8 shows a generalized depiction of the Globalstar satellite architecture. Immediately obvious is the simplicity of the design. There is no onboard processing here; all of the technology is low risk. A hallmark of the Globalstar system is that the complexity, where needed, is relegated to the ground segment.

The Globalstar system is active at the time of this writing (2006). Commercially the system has not been successful, at least for the original investors. In contrast to Iridium, there are a number of factors about Globalstar that make its value proposition quite different. Some of these factors include

- *Simple architecture*: The space segment is constructed with low-risk technology. The complexity of the system is kept on the ground where upgrades, as necessary, are readily accomplished. This approach keeps the overall cost of the system down, and permits lower service pricing to

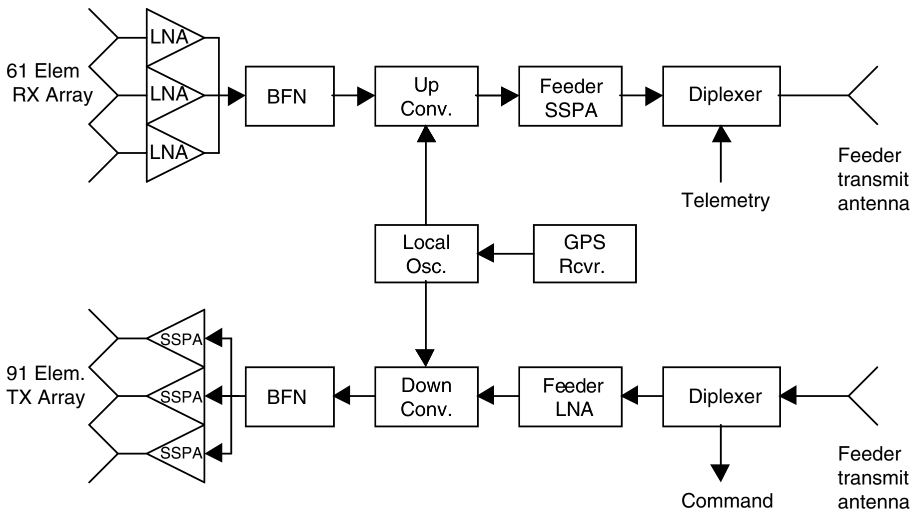


FIGURE 11.8 General architecture of the Globalstar satellite.

be offered. In fact, the start of service was offered at $1.79 per minute, and a handset price of around $1500 [23], a far cry from Iridium's $7 per minute and a $3000 handset. Currently, Globalstar services can be had for about $1.50 per minute, and the handset costs about $750.

- *Proven air interface*: The CDMA-based IS-95 air interface has been well proven in terrestrial cellular systems. In addition, this approach makes the job of constructing multimode UTs (terrestrial-satellite) easier.
- *Phased service rollout*: Globalstar observed the techniques employed by its Iridium forerunner and decided to gradually roll out service in well-researched market areas. This technique allowed the company to gain experience and make corrections as needed with a fault-tolerant, methodical process.
- *Revenue area focus*: Globalstar is not attempting to service low population areas like the oceans or the polar regions. The focus is on the parts of the globe where the people are. Precious resources are not expended in areas where return is marginal.
- *More esthetically-pleasing UT*: Although not quite the size of a modern-day cellular handset, and the antenna is still quite a bit larger than desirable, the Globalstar UT is much smaller than its Iridium counterpart. The size is still too large for broad user acceptance.

Globalstar is now largely focused on the same niche markets that Iridium is targeting, although the system does not have coverage at the Poles. This marketing approach can be seen by virtue of a statement recently seen on their website

> Globalstar voice and data customers include businesses who operate in areas where cellular coverage is poor or non-existent and landline service unavailable. Natural resource companies, long-haul transportation operators, commercial fishermen, government employees, recreational and travel enterprises, geologists, prospectors and public safety organizations all value Globalstar products and services. Outdoor enthusiasts can have the convenience and safety of a Globalstar handset while they fish, hunt and enjoy themselves in remote areas.

11.4.2 MEO

MEO systems, in general, tend to be orbitally located in the region between the inner and outer Van Allen radiation belts (around 10,300 km). Consequently, they tend to take on a blend of characteristics of which some are LEO-like and others are GEO-like. For instance, MEO propagation delay works out to be around

40–50 ms, not as good as LEO, but clearly better than GEO. The satellites also tend to be fairly complex and approach the GEOs in terms of design life (on the order of 10–12 years). Because of their relatively high orbit, MEO satellites move fairly slowly across the sky, greatly simplifying tracking requirements and reducing the number of hand-offs during a typical call holding period.

Although a number of MEO-based cellular satellite systems have been proposed in recent years, none has actually been fielded in any significant sense (i.e., commercially operational). As examples of systems that have been considered, we will examine two of the better-known MEO systems here: ICO and Ellipso.

11.4.2.1 ICO [15,24]

The design of the ICO system began with a study program conducted by Inmarsat in the early part of the 1990s. Dubbed "Project 21," Inmarsat's objective was to move into the satellite cellular communications business. LEO, MEO, and GEO constellations were considered during the study, which eventually settled on a MEO configuration (an "intermediate circular orbit," or ICO), subsequent to which the project was spun off as a separate company. ICO was owned by ICO Global Communications Holdings, Ltd., with an additional 17 subsidiaries. That was until the company sought bankruptcy protection around the same time as Iridium faced a similar difficulty. Then [25], Craig McCaw and his Eagle River organization, along with Subash Chandra of ASC Enterprises (Ascel), both saw potential in the system and, essentially, took over the company. The new owner organization was known as New Satco Holdings, Inc. The current company is owned by ICO Global Communications (Holdings) Ltd., incorporated in Delaware, USA. The major equipment contractors for the system included Hughes Space and Communications (satellites, now Boeing) and a team, lead by NEC, which included Ericsson and Hughes Network Systems (ground infrastructure).

The original focus of the ICO system was to complement terrestrial cellular communication by servicing customers who reside outside of normal cellular coverage, providing cellular extension service to those who often travel outside of terrestrial cellular, maritime customers, and also government users. In other words, ICO was already targeting a fairly broad market. Craig McCaw, on the other hand, saw opportunities to enhance the system to provide data services. There were indications that the design was being altered to include packet-based medium data rate services (GPRS-like rates of around 384 kb/s), with early support for Wireless Application Protocol (WAP) services. WAP enables thin clients, like cell phones, to access enterprise services like email and Internet information.

The full ICO constellation was planned to consist of 10 active satellites in two planes of five satellites each. The orbital planes are both circular and inclined at 45°. The satellites communicate with 12 gateways (ground stations), called satellite access nodes (SANs), spread around the world. Each SAN (see Figure 11.9) is equipped with five antennas to be able to track as many satellites as might be in view at any given time.

Owing to the higher altitude of the constellation, each satellite is able to cover around 25% of the Earth's surface. The air interface is very similar to the terrestrial cellular IS-136 (D-AMPS) standard, in that it is an FDMA/TDMA scheme (QPSK modulation) that can support as many as 40 timeslots (nominal-rate users) on five subcarriers within a 156 kHz channel bandwidth. The fact that the channel bandwidth is reasonably wide (approximately 156 kHz) is the reason that the system is readily modified to accept GPRS-like (and even EDGE-like) waveforms. There is, therefore, scope to provide packet-based medium data rate services in the ICO system, and this is one of the "future proofing" characteristics that were designed in.

Mobile downlinks are at L-band (1980–2010 MHz) while the uplinks are at S-band (2170–2200 MHz). The aggregated signals, shipped between the satellites and the SANs, are transmitted at C-band (5187–5237 MHz up, and 7018–7068 MHz down).

The ICO Network Management Center (NMC) is located in Tokyo, Japan, while the Satellite Control Center (SCC) has been placed in Uxbridge, England. The SCC monitors the health and status of the spacecraft, and also takes care of any orbital adjustments that might become necessary in the course of the mission life.

The ICO spacecraft design is fairly sophisticated. Moderately large (2600 kg and consuming around 8700 W), the satellite is based on the HS-601 design, which has been extensively deployed for

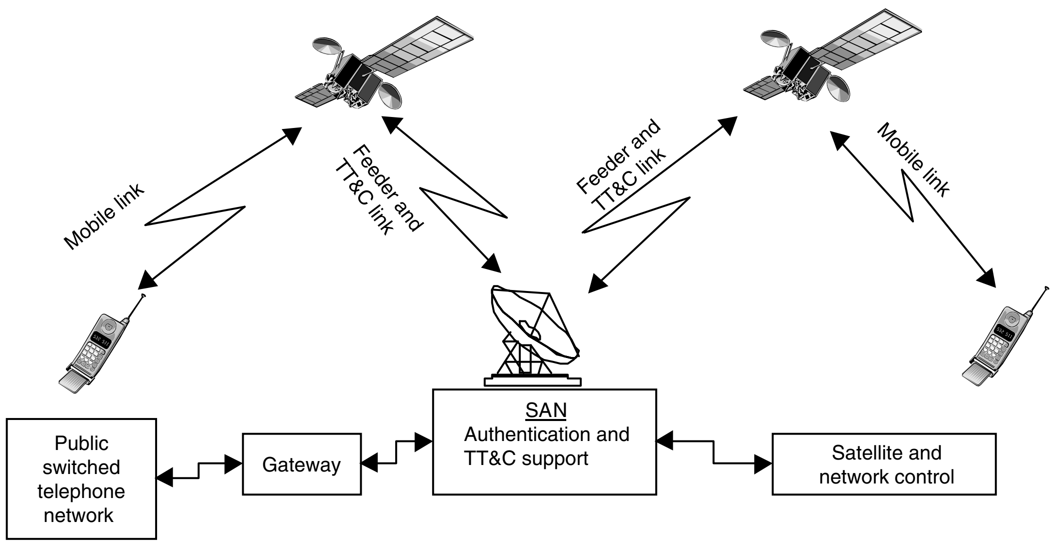


FIGURE 11.9 Generalized ICO system architecture.

GEO missions. There are two mobile link antennas, one for transmit and the other for receive, each measuring around 2 m in diameter. These antennas are direct radiating arrays driven by a sophisticated DBF, which allows the antenna patterns (cells) to be dynamically shaped in order to respond to changing loading needs. This design also means that complexity (cost) has been added to the spacecraft, and that it must be supported by an intricate calibration infrastructure as well. Each antenna subsystem consists of 127 radiating elements and forms around 163 beams that, together, cover a ground surface diameter of around 12,900 km. The system frequency reuse factor is about four times with this design. Each satellite has switching capacity for around 6000 voice channels, although power constraints limit the actual capacity to around 4500 circuits. On the other hand, given the system modifications noted earlier, the system is poised to enter the packet switched regime. Entirely different capacity calculations are possible under such an operation environment.

ICO's development has been a mixed bag, and apart from McCaw and Chandra's involvement, the value proposition is uncertain. Factors include

- *Higher orbit, slower movement, fewer hand-offs*: This characteristic makes for easier tracking and allows certain simplifications in the ground infrastructure.
- *Smaller constellation (with respect to LEO)*: Fewer satellites to build and launch, but each satellite is heavier and more complex.
- *Wider field-of-view*: Higher orbit sees more of the earth, and can cover oceanic regions (a mixed blessing).
- *Fewer gateways required*: Lower infrastructure cost than LEO, but higher cost than a GEO system.
- *IS-136 (D-AMPS)-like air interface*: Allows adaptation of standard terrestrial cellular hardware and easier manufacture of multimode handsets.

ICO has clearly had a tough time getting started, and for a time it appeared that the system would suffer the same fate as Iridium. The system is relatively expensive (included a significant amount of innovative technology). Again, estimates vary, but the overall system cost around $3 to $5 billion in its original state. Packet-based modifications will add additional cost. With the further innovations to wireless packet data support, the future of the ICO system will be interesting to watch as it unfolds.

11.4.2.2 Ellipso [15,26,27]

Ellipso has also been referred to as being in the class of the "Big LEOs." The system orbit design is unique, to the point of actually having been patented (U.S. Patent 5,582,367). Owing to the altitude (mean altitude, in the case of the elliptical planes) we class the Ellipso system here as a MEO system.

Mobile Communications Holdings, Inc. owns Ellipso. Initially, the major contractors were Lockheed Martin (ground) and Harris (space) supplying equipment to the system. Services are planned to be provided globally, though biased to favor populated areas, through satellites in three orbital planes; two elliptical ones that are called "Borealis" and one circular called "Concordia." The orbits are optimized to provide regional coverage proportional to the distribution of population on the surface of the earth. The Borealis orbits are elliptical, sun-synchronous, inclined at 116° and each contain five satellites. These orbits each have a perigee at 520 km and an apogee at 7846 km. The Concordia orbit, on the other hand, is equatorial, circular and has seven equally spaced satellites. Concordia's altitude is 8060 km. Both of these orbits are well within the band separating the two Van Allen radiation belts.

Ellipso's guiding philosophy is to perform all system trades with an eye toward the lowest end-cost to the subscriber. Its services (including voice, messaging, positioning, and Internet access) are targeted toward "everyman." In other words, in contrast to the target customers of Iridium and ICO, Ellipso wants to reach deeper into the market and not just focus on the affluent, globetrotting businessman.

Like the ICO system, Ellipso provides a mobile user downlink at L-band (1610–1621.35 MHz) with the uplink placed at S-band (2,483.5–2,500 MHz). Feeder (ground station) communication, on the other hand, is performed with the uplink at Ku-band (15,450–15,650 MHz) and the downlink at high C-band, nearly X-band (6,875–7,075 MHz). Each satellite forms a cell beam pattern consisting of 61 spot beams incorporating a high degree of frequency reuse (by way of cell isolation and orthogonal coding) across the coverage pattern. The system air interface is based on third-generation (3G) wideband CDMA (W-CDMA) with an occupied bandwidth of 5 MHz. This is a technology that is only just starting to be deployed in the terrestrial cellular world at the time of this writing. Consequently, the Ellipso design exhibits a tremendous degree of forethought and "future-proof" planning. Because of its W-CDMA infrastructure, Ellipso is poised to launch services with a 3G infrastructure in place and is, consequently, ready to provide high data rate (up to 2 Mb/s) packet-based services.

A general diagram of the overall Ellipso system is shown in Figure 11.10.

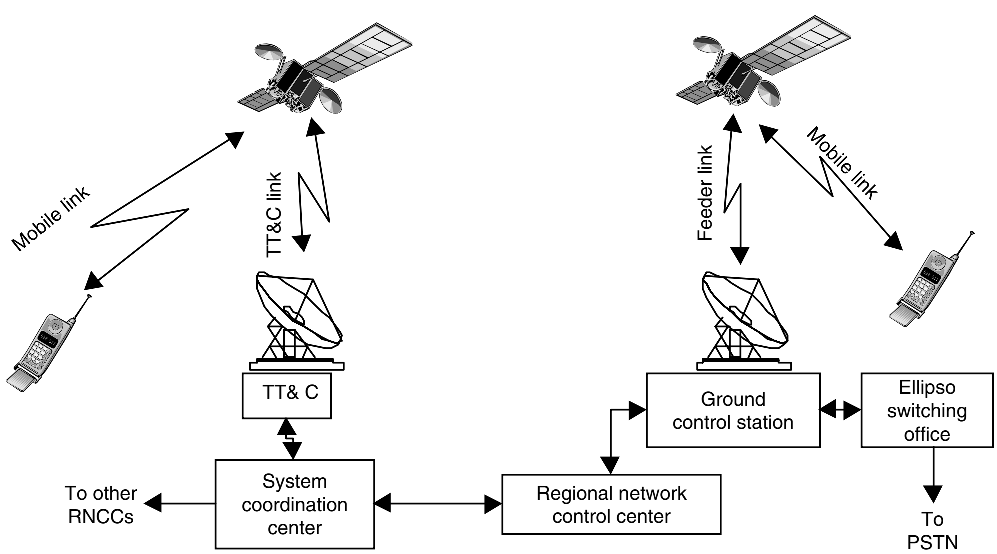


FIGURE 11.10 General diagram of the Ellipso MEO system.

The central System Coordination Center takes care of the system network planning and monitoring. The Ground Control Stations (GCS) provide the gateway function as the interface to the communication signals going to, and coming from, the mobile stations via the satellites. Regional Network Control Stations provide local network control functions and collection of billing records. Associated with each GCS is an Ellipso Switching Office (ESO), which provides the interface between the PSTN and the Ellipso system. The ESO, in addition, houses the HLR and VLR, and takes care of the user authentication function.

All Ellipso satellites are of identical design, regardless of the orbit into which they are placed. Simplicity is the driver behind the satellite design as well. They are straightforward bent pipe translators with separate transmit and receive, direct radiating, fixed beamformed phased array antennas with 127 radiating elements in each planar array. Each satellite is of medium size with a mass of around 700 kg which, in turn, keeps launch costs down.

The success of Ellipso has yet to be seen at this juncture. The system does, however, contain many of the features important to an attractive value proposition. This is another system that will be interesting to watch as development unfolds. Some of Ellipso's more interesting value features include

- *Low system cost*: The system designers are fanatical about keeping costs down as they are keenly aware of the connection between system deployment cost and the ultimate price point that can be offered to the subscriber.
- *Wideband 3G-compatible air interface*: A very forward-looking design feature. The system will be ready for enhanced 3G services, including always connected, high-speed packet-based Internet access. This feature has the potential of being very attractive to potential subscribers.
- *Complexity on the ground*: The space segment is designed as simply as practical in the overall system context. Complexity is kept on the ground where system upgrades are more readily accomplished.
- *Low price point and large target market*: The designers are targeting average consumers and not focusing on the lower quantity of affluent business travelers.
- *Revenue area concentration*: System design focuses on areas on the earth where the bulk of the people are. Resources are not wasted over large ocean regions and areas of sparse population.
- *Phased deployment*: The characteristics of the Ellipso orbits are such that service can be initiated with a partially deployed constellation of satellites. In effect, early revenues can be generated that will help to pay for the remainder of the system deployment.

11.4.3 GEO

Two of the most advanced GEO-based cellular satellite systems are the ACeS and the Thuraya systems. Both of these systems were put into commercial service in 2000.

11.4.3.1 ACeS [15,28]

ACeS (Asia Cellular Satellite) is a GEO cellular satellite system conceived, designed, and backed by a partnership consisting of P.T. Pasifik Satelit Nusantara (of Indonesia), the Philippine Long Distance Telephone Company (PLDT), Lockheed Martin Global Telecommunications (LMGT), and Jasmine International Overseas Company, Ltd. (of Thailand).

The ACeS primary target market is the Southeast Asian area comprising the 5000 islands of Indonesia in the south, Northern China in the north, Pakistan in the west, and Japan in the east (see Figure 11.11).

The coverage area encompasses some three billion people, many of who have little or no access to a wired communication infrastructure. The Indonesian archipelago, for instance, is an expanse of islands that stretches some 4000 miles from east to west. It is not difficult to imagine the tremendous challenge of building a wired infrastructure to interconnect such a country. As such, the ACeS service is focused on a tightly defined market region, widely recognized as a rapidly expanding industrial world sector. While the typical ACeS user is perceived to be an active business traveler, the pricing of the planned services is expected to be at a level well within the reach of middle-class business people. Consequently, ACeS has a large addressable user population. This approach stands in stark contrast to some systems that target the high-end traveling businessman and seek to provide complete global coverage. In itself, a global

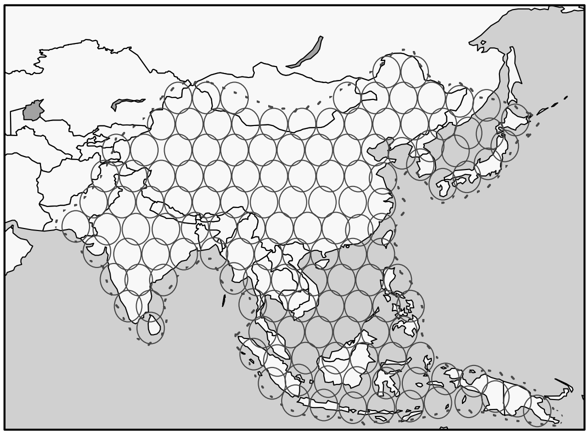


FIGURE 11.11 ACeS coverage and beam.

FIGURE 11.12 ACeS handset.

coverage system, like some low earth orbiting (LEO) systems, requires a large quantity of satellites (along with sophisticated hand-off and traffic management methods) and has an attendant high implementation price tag.

ACeS is designed to operate in a clearly defined geographical coverage area, therefore a regional GEO-based satellite system was chosen. Such a well-defined coverage area ensures that a maximum amount of precious satellite resources is concentrated on the desired revenue-producing areas. The satellite air interface standard was based on the ubiquitous GSM terrestrial cellular standard in order to take advantage of its feature-rich suite of services, as well as the availability of a large quantity of standard supporting hardware. A GSM-based system also eases the integration of terrestrial hardware (e.g., dual-mode ACeS-GSM handsets) and facilitates intersystem roaming (on the basis of GSM Subscriber Identity Module [SIM] cards). A mutually beneficial cooperative effort between LMGT and Ericsson (a world-class supplier of mobile communication equipment) assured the definition of an optimally tailored AIS. Further, the earth-satellite links were conservatively designed to ensure good service to disadvantaged users; frequently dropped calls are death to service acceptance and customer loyalty. An essential component of customer acceptance, facilitated by the strong links, is a handset form factor that is comparable to what is expected in modern cellular handsets. Figure 11.12 shows a picture of the ACeS handset in the form factor to be used at service launch.

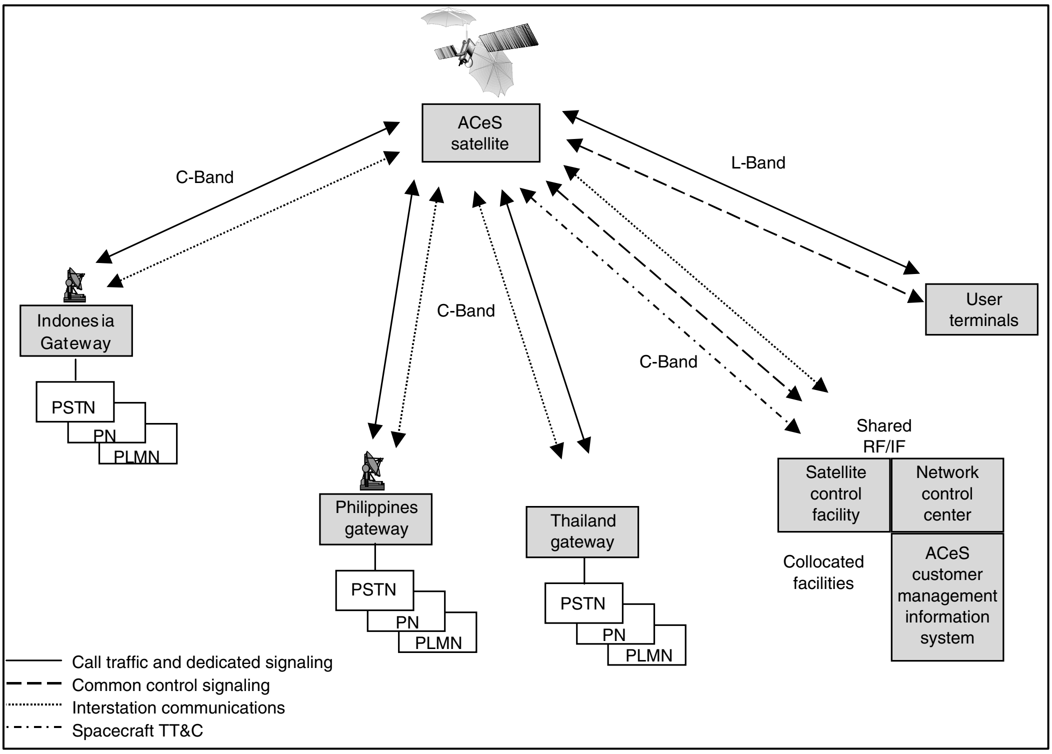


FIGURE 11.13 The ACeS system functions.

In addition, low cost of both service (airtime) and equipment is essential to customer uptake and heavy system use. Again, the modest deployment cost of a regional GEO system implementation aids in keeping the costs low.

The ACeS system has two main components, namely, the ground segment and the space segment, where the ground segment is further subdivided into the backhaul and control function (implemented at C-band) and the L-band user link function (see Figure 11.13).

There is one satellite control facility (SCF) for each spacecraft. The NCC provides the overall control and management of the ACeS system, including such functions as resource management, call setup and tear down, call detail records, and billing support (customer management information system). Regional gateways, operated by the various National Service Providers (NSPs) manage the subset of system resources as allocated by the NCC. These gateways also provide the local interface and billing to the actual system users, and also provide connectivity between ACeS users and the wired infrastructure (Public Switched Telephone Network, Private Networks, and/or the Public Land Mobile Network). The user segment consists of handheld mobile or fixed terminals. These terminals can be configured to provide basic digital voice, data, and fax services. Further, since the system is based on GSM at the physical layer (in particular, 200 kHz Time Division Multiple Access, or TDMA, channels), it is future proofed in the sense that it will support GPRS and Enhanced Data for GSM Evolution (EDGE) upgrades with *no change required to the space segment.* This feature is a very important characteristic of the ACeS system.

The ACeS spacecraft, dubbed Garuda-1, is one of a family of Lockheed Martin modular spacecraft in the A2100-series (see Figure 11.14).

In particular, Garuda-1 is, an A2100AXX, one of the largest models. This spacecraft is three-axis stabilized, with a bus subsystem that has been applied across numerous other spacecraft programs (e.g., Echostar, GE, LMI, and others) and, consequently, has a significant amount of application heritage. The payload is designed with two separate L-band antennas on the user link side. One antenna is dedicated to

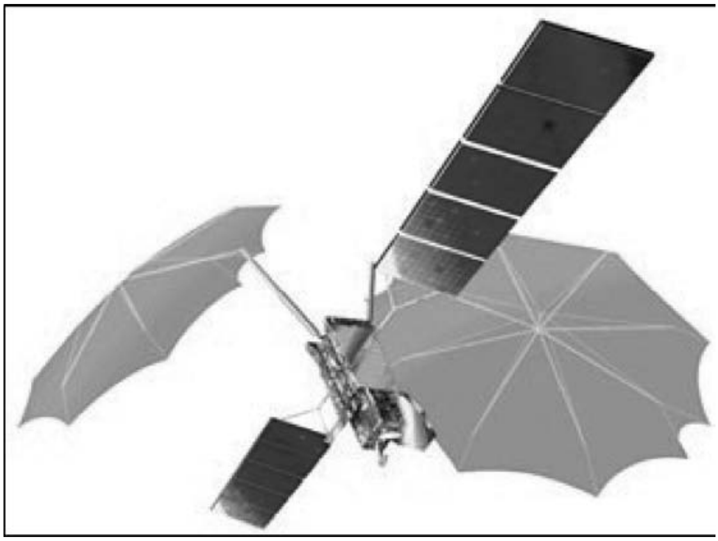


FIGURE 11.14 ACeS Spacecraft: deployed.

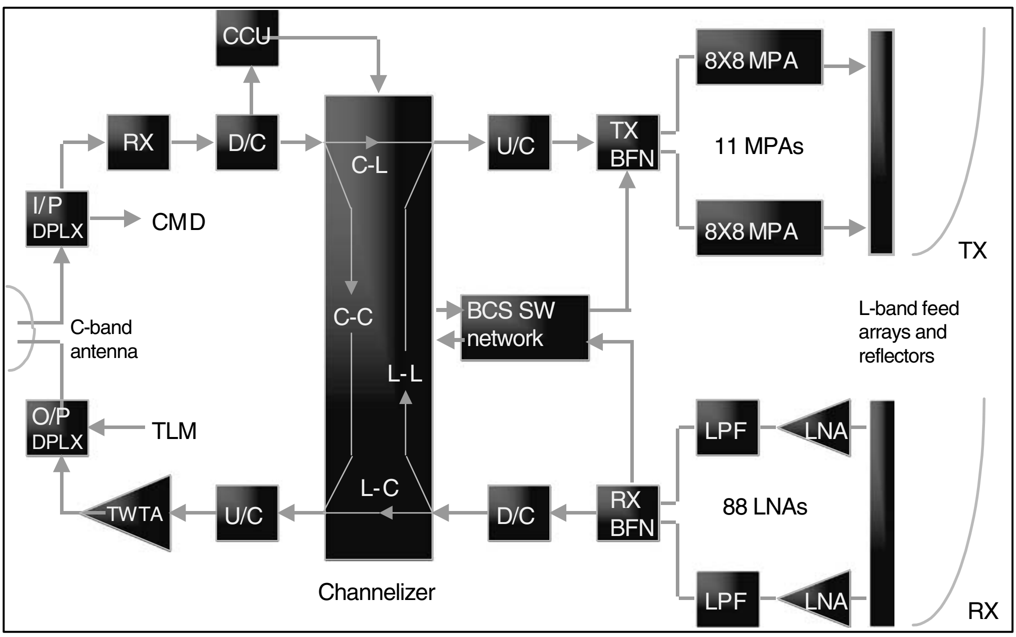


FIGURE 11.15 ACeS payload block diagram.

the transmit function (Forward link) with the other, naturally, for receive (return link). This separation of antenna functions was carried out to minimize the probability of receive-side interference from passive intermodulation (PIM) in the transmit side. Large (12 m) projected antenna apertures (two of them) provide the underpinnings for strong user links, crucial to reliable call completion under a variety of user circumstances. The result is high aggregate EIRP, at 73 dBW, and high-G/T, at +15.3 dB/K.

Figure 11.15 illustrates the major block functions of the Garuda-1 communication subsystem (CSS).

As noted earlier, the user link is closed in the conventional mobile satellite system L-band. A total of 140 narrow beams are formed with a low-risk, low-power analog beamforming network (BFN) approach, in both forward and return directions. This beam design gives rise to a composite pattern that provides 20 times frequency reuse, thereby efficiently conserving valuable spectrum. A beam congruency system (BCS) operates in conjunction with ground beacons to ensure proper overlap of the corresponding transmit and receive beams. The L-band transmit amplification subsystem is implemented with a distributed set of Butler matrix based amplifier blocks called matrix power amplifiers (MPAs). The MPA construct provides equal loading for all amplifiers in the block, which, in turn, minimizes phase and amplitude variation across the block. Small variations are key to good isolation between beams, giving good frequency reuse performance and, hence, maximum traffic capacity.

At the heart of the CSS sits the digital channelizer. This channelizer performs the important function of filtering the TDMA traffic channels and, in addition, of routing the individual traffic bursts as needed (particularly for the case of direct mobile-to-mobile communication).

The balance of the CSS (shown to the left of the channelizer) is the C-band backhaul transmit and receive equipment. This equipment is, largely, heritage being very similar to that used on previous direct broadcast and fixed service satellites.

The ACeS system has several important characteristics that hallmark a successful system. Among these characteristics are

- *Low infrastructure cost*: Being a GEO system, once the satellite is on-station, instant connectivity is possible. The entire system, both space and ground segments, are reported to be under $1 billion (about an order of magnitude less than most Big LEO systems).
- *Low targeted service cost*: Low infrastructure cost allows lower price points to be charged while still allowing the business to turn a profit. This approach also means that a greater market can be attracted for the service.
- *Well-researched target market*: The Asia-Pacific region includes a multitude of islands that are not well connected by any kind of terrestrial infrastructure. This area is also one of high industrial growth which, in turn, needs good communication for support.
- *Esthetically-pleasing UT*: The satellite bears the burden of providing the link margin, so that the UT can be sized similar to terrestrial cellular telephones. On the downside, complexity is added to the space segment thereby adding cost to the satellite. Since only one satellite is involved, this added complexity is easier to bear.
- *Well-chosen air interface*: The ACeS air interface is based around the GSM standard, implying the ability to directly use terrestrial hardware in the terrestrial equipment (thereby reducing development cost).
  - *Wideband*: 200 kHz channels (the GSM standard) will support future 2.5G services (GPRS and EDGE) without modification to the space segment. This is a design forethought that helps to future-proof the system.
  - *Adjacent service compatibility*: The adjacent terrestrial cellular services are largely based on GSM. Consequently, multimode UTs can be provided in a cost-effective manner.

11.4.3.2 Thuraya [15,29,30]

A partnership group led by Etisalat (Emirates Telecommunication Corporation) of the United Arab Emirates owns the Thuraya system. The objective of the system is to provide regional cellular satellite service to an area that includes Continental Europe, Northern Africa, the Middle East, and India. The Thuraya coverage area is adjacent to the ACeS coverage area and, in addition, shares many of the same design features seen in that system. The target subscriber population includes national and regional roamers in an area (desert) that is not well served by terrestrial cellular service. Types of services offered include voice, facsimile, low-rate data, short messaging, and position determination. Major suppliers of the system components included Hughes Space and Communications (space segment), Hughes Network Systems (ground infrastructure), and Ericsson (network switching equipment).

The Thuraya system has one GEO satellite positioned at 44°-East longitude (a second satellite has also been launched and placed at 28.5°-East. Mobile users connect with the Thuraya system through handheld

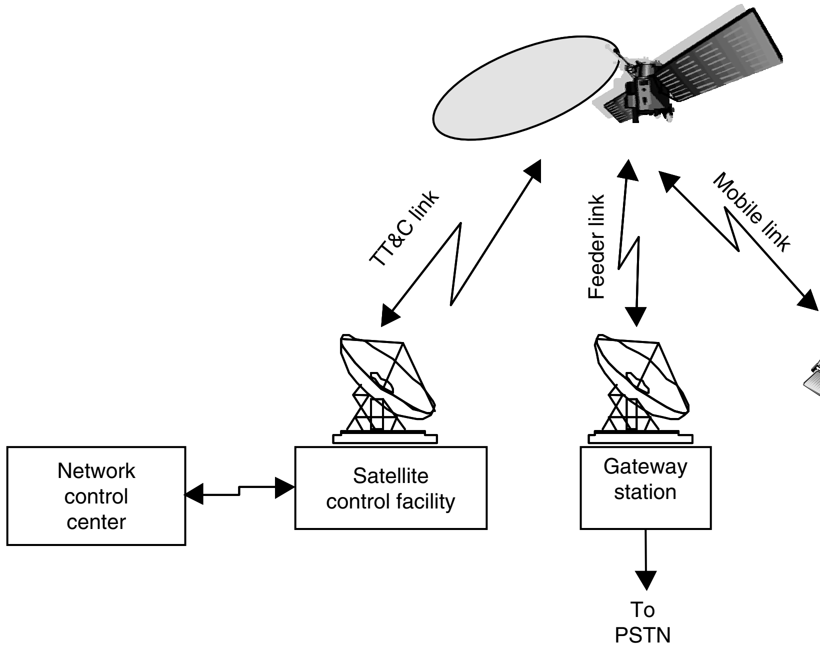


FIGURE 11.16 General diagram of the Thuraya system.

UTs operating at L-band (1626.5–1660.5 MHz uplink and 1525–1559 MHz downlink). The aggregated signals destined for connection to the public wired infrastructure (PSTNs, etc.) are connected between the satellite and the gateway stations at C-band (6425–6725 MHz up and 3400–3625 MHz down). The air interface is similar to that developed by Hughes Network Systems for the ICO system. That is, it is similar to the terrestrial cellular IS-136 (D-AMPS) system with Offset QPSK modulation, although the higher protocol elements are GSM compatible. Consequently, the Thuraya system is GSM compatible on a network and service level. Figure 11.16 shows the main elements that comprise the system.

The Gateway station serves as the main interface for communication signals into the public infrastructure. The system has a primary gateway that is located in Sharjah, UAE. Accommodation for several regional gateway stations is provided in the network infrastructure. The main functions performed by the gateway include user authentication (HLR and VLR), call control (set up and tear down), billing records, resource allocation, and roaming support. All interface conditioning required to connect to PSTNs, PLMNs, and so on are also handled in the gateway.

The SCF is actually classed as part of the space segment in that it handles the monitoring and control of the satellite. All telemetry is monitored in the SCF, and the required bus and payload commands are initiated at this facility. The NCC, as the name implies, handles all of the network administration functions (routing, congestion control, and similar functions). Any payload commands required in this context are translated by the SCF and sent to the satellite.

The satellite is a very sophisticated element of the Thuraya system. It is a very large spacecraft (4500 kg) with a mission life of 12 years. The spacecraft is based on the well-known HS-601, but with several enhancements. For instance, solar concentrators are used on the solar arrays in order to enhance collection. This modification is performed in order to help supply the large quantity of electrical power required (12.5 kW). The satellite is capable of switching about 13,750 simultaneous voice circuits. Mobile links interface to the ground through a single 12.25-m projected aperture reflector. PIM, that is, unwanted mixing noise from the transmitter entering the receiver, is normally a central concern in systems such as these. Mitigation methods in other systems (e.g., Inmarsat-3 and ACeS) have included the use of two separate antenna apertures. The engineering of the Thuraya system has solved that problem, and only

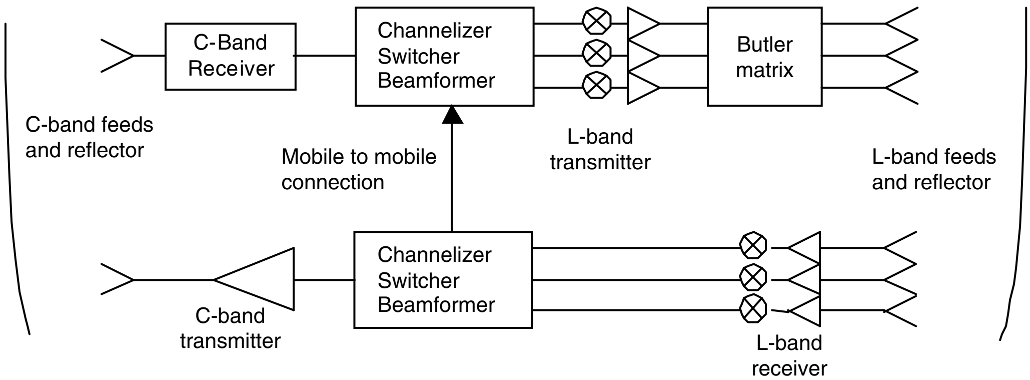


FIGURE 11.17 Major elements of the Thuraya communications payload.

one, large reflector is required (serving both transmit and receive functions). The 256 cellular beams are generated by a state-of-the-art DBF, which allows dynamic beamshaping as required to meet variations in traffic loading, as required. This DBF, though adding a great deal of flexibility to the system, comes at a price of added complexity (cost), power consumption, and a necessary dynamic calibration infrastructure. Some of the blocks of the CSS of the Thuraya satellite are shown in Figure 11.17.

As a GEO system, Thuraya has many of the same positive value characteristics described previously:

- *Low infrastructure cost*: Thuraya has the instant connectivity trait common to GEO systems where only one satellite is required. The system cost is reported to be around $1 billion.
- *Low targeted service cost*: Low price points are planned to be for system services (reportedly, on the order of $0.50 per minute). This approach also means that a greater market can be attracted for the service.
- *Target market*: The northern Africa area contains a lot of open territory (desert) where a significant industry is conducted (e.g., petroleum). Existing terrestrial infrastructure is, clearly, not adequate so a good business opportunity exists. On the other hand, the potential for generating significant revenue from European coverage area is questionable given the ubiquity of GSM in that region.
- *Esthetically-pleasing UT*: Thuraya, with its 12.25-m antenna aperture, allows the UT can be sized similar to terrestrial cellular telephones. On the downside, complexity and cost accrues to the satellite as a result. As we noted in the ACeS case, only one satellite is involved so this added complexity is easier to bear.
- *Air interface*: Thuraya's air interface shares many of the characteristics of the D-AMPS IS-136 standard, implying the ability to directly use terrestrial hardware in the terrestrial equipment (thereby reducing development cost). It is also relatively wideband (around 156 kHz) and is, therefore, suited to support 2.5G services as they become available.
- *GSM network infrastructure*: The network services are based on the GSM model and can support GSM services as a result.

11.5 Trends

Satellite cellular service developments have, by and large, mimicked the advancements in terrestrial cellular [31–33], albeit with a predictable delay. Clearly, the satellite service infrastructure takes longer to develop and field than its terrestrial counterpart. Cellular satellite service providers have continued to look to the terrestrial developments for the "next step." Activities are under way to adapt existing systems, and to incorporate future enhancements, to effectively support data transmission. The immense data communication infrastructure demands seen today are being fueled (akin to gasoline being fire-hosed onto a blaze!) by the explosive growth of the Internet. Cellular satellite system developers aim to have a part in the on-going explosion. Internet traffic, from the user's perspective, is primarily a bursty form of communication based on packet switching. Circuit switched systems are not efficient in bursty packet mode, so cellular satellite systems are being adapted to communicate in packet mode and at higher data rates. Mobile data communications require packet data access on the order of 100 kb/s or higher. GPRS, in GSM, accommodates dynamically adjustable rates and, in eight-slot full-rate mode will support a peak rate on the order of 115 kb/s. If EDGE is included (with adaptive coding and its 8PSK modulation approach), it will squeeze 384 kb/s into a 200 kHz GSM channel. Plainly, cellular satellite systems based on GSM will readily be able to support GPRS and EDGE.

3G W-CDMA terrestrial systems are being deployed in order to support mobile data rates as high as 2 Mb/s. One cellular satellite system design (Ellipso) has this concept in mind. Others may follow. Teledesic, for instance, is a wideband system concept that at its inception was driven by Bill Gates and Craig McCaw. This system was conceived as a LEO system specifically designed to provide high-speed data to the home fixed locations), but it is not a cellular satellite system under the criteria we have used here.

Other terrestrial enhancements are in the works, and it is likely that their incorporation into the cellular satellite world will occur. For instance, the Wireless Application Protocol (WAP) is a protocol stack specifically designed to allow "thin clients" (limited capability devices like cellular telephones) to take greater advantage of the Internet. WAP allows a more streamlined approach for cell phones to receive and transmit email, browse Internet web sites, interact with corporate enterprise structures (Intranet services), and other well-established Internet-based activities.

Cellular satellite systems have most successful in specific applications where satellites do the best job [34,35], namely

- Access to remote places not adequately served
- Extension of service coverage over a wide region, such as a large country, continent or ocean; especially if large populations are involved (e.g., Southeast Asia, Indonesia, etc.)
- Extending services across borders: achieving a trans-national network
- Delivery of communication services, particularly emergency services, when terrestrial networks break or lose power (e.g., natural disasters)

The need for cellular satellite systems has yet to run its course. Applications targeting areas that are the domain of terrestrial cellular are in the works. For instance, a recent FCC ruling allows the use of the so-called ancillary terrestrial component, which allows radio resources to be shared between the satellite and the terrestrial segments of a satellite cellular communication system. One company, Mobile Satellite Ventures LP, is already applying this methodology [36]. It would seem that the future of satellite cellular systems is one of complementary function with its terrestrial counterpart, bringing a more complete palette of communication services worldwide.

References

1. Defense Department Contract Keeps Iridium Satellites in Orbit, http://www.iridium.com/corp/iri_corp-news.asp?newsid=14
2. Inmarsat Web site, http://about.inmarsat.com/satellites.aspx?top_level_id=3& language=EN& textonly=False
3. Quality of Service Forum web page, http://www.QoSforum.com
4. Worldwide B2B e-Commerce market, http://www.epaynews.com/statistics/bankstats.html (Gartner Group).
5. eGlobal Report, March 2000, http://www.emarketer.com; this report predicted the worldwide e-Commerce market to grow to $1.44 trillion, and actually hit $3.6 trillion (Gartner).
6. Yates, D., "M-Commerce, twice the cash value of e-commerce," NetRegistry News Centre, September, 2005, http://www.netregistry.com.au/news/articles/4/1/M-Commerce,-twice-the-cash-value-of-e-commerce
7. Butash, T.C., *BAE Systems*, Manassas, private communication, 2000.

8. Matolak, D., *Lockheed Martin Global Telecommunications*, private communication, 1998.
9. Kadowaki, N., et al., ATM transmission performance over the trans-Pacific HDR satcom link, *Proc. Second International Workshop on Satellite Communications in the Global Information Infrastructure*, 63, 1997.
10. Fitch, M., ATM over satellite, *Proc. Second International Workshop on Satellite Communications in the Global Information Infrastructure*, 67, 1997.
11. Falk, A., TCP over satellite, *Proc. Second International Workshop on Satellite Communications in the Global Information Infrastructure*, 74, 1997.
12. Johnson, Lt. G.W. and Wiggins, ET1 M.D., Improved Coast Guard communications using commercial satellites and WWW technology, *Proc. Fifth International Mobile Satellite Conference*, 519, 1997.
13. Logsdon, T., *Mobile Communication Satellites*, McGraw-Hill, Inc., New York, 1995, 131.
14. Stern, P. and Mauricio, P. The exploration of the Earth's magnetosphere, *NASA Web Tutorial*, 1998, http://www-spof.gsfc.nasa.gov/Education
15. Miller, B., Satellites free the mobile phone, *IEEE Spectrum*, 26, March 1998.
16. Elizondo, E. et al., Success criteria for the next generation of space based multimedia systems, *Satellite Communications Symposium of the International Astronautical Federation*, Session 33, September 1998.
17. Maine, K., et al., Overview of Iridium satellite network, *Motorola Satellite Communications Division*, Chandler, AZ, 1995.
18. Lloyd's Satellite Constellations, *Big LEO tables*, web link: http://www.ee.surrey.ac.uk/Personal/L.Wood/constellations/tables/
19. Iridium satellite system purchased by new firm, Spaceflight Now, November, 2000, http://www.spaceflightnow.com/news/n0011/16iridium/
20. The author has personally spoken with a number of engineers who worked on the Iridium system. Without exception, every one of these engineers had glowing memories of the technical challenges and immense job satisfaction that came from being a part of the engineering team.
21. Dietrich, F.J., The Globalstar satellite cellular communication system design and status, *Proc. Fifth International Mobile Satellite Conference*, 139, 1997.
22. History of the Handheld Satellite Phone, Globalcom, 2006, http://www.globalcomsatphone.com/articles/history.html
23. Wilkinson, G., Satellite companies face murky fate after Iridium's demise, *Total Telecom*, March 24, 2000.
24. ICO Communications, web site: http://www.ico.com/
25. Foley, T., Mobile & satellite: McCaw unveils WAP strategy for new ICO, *Communications Week International*, April 3, 2000.
26. Draim, J.E., et al., Ellipso—An affordable global, mobile personal communications system, *Proc. Fifth International Mobile Satellite Conference*, 153, 1997.
27. Ellipso web site: http://www.ellipso.com
28. Nguyen, N.P., et al., The Asia cellular satellite system, *Proc. Fifth International Mobile Satellite Conference*, 145, 1997.
29. Alexovich, A., et al., The Hughes geo-mobile satellite system, *Proc. Fifth International Mobile Satellite Conference*, 159, 1997.
30. Thuraya web site: http://www.thuraya.com
31. The future: Down the road for PCS computing, *PCS Data Knowledge Site*, Intel, 1998, http://www.pcsdata.com/future.htm
32. Route to W-CDMA, *Ericsson Mobile Systems*, Ericsson, December, 1998, http://www.ericsson.com/wireless/products
33. Cadwalader, D., Toward global IMT-2000, *Ericsson Wireless NOW!*, January 1998, http://www.ericsson.com/WN/wn1-98/imt2000.html
34. Greenfield, H., "Satellites and IT: At the Digital Crossroads," *Satmagazine*, January, 2006, p. 19ff.
35. Elbert, B., Satellites and IT—Myths and Realities, *Satmagazine*, January, 2006, p. 33.
36. Mobile Satellite Ventures Examines Advantages of New Communications Technology in Wake of Hurricane Katrina, *Space Newsfeed*, October, 2005, http://www.spacenewsfeed.co.uk/2005/30October2005_3.html

12

Electronics for Defense Applications

Michael J. Biercuk
Booz Allen Hamilton/Defense Advanced Research Projects Agency

John C. Zolper
Defense Advanced Research Projects Agency

12.1 Introduction

Radio frequency (RF) and microwave systems have played a significant role in service of defense objectives since the dawn of the electronics era. Among the earliest applications of RF electronic equipment for defense was the area of communications. Battlefield dominance and strategic victories have historically relied on the ability to communicate effectively over long distances in real-time, as exemplified by the courageous story of Cher Ami, a carrier pigeon flown by the U.S. Army Signal Corps in Verdun France during World War I [1]. In that same conflict niche communications applications appeared for radio, as air–ground radio communications facilitated reconnaissance missions flown by military pilots [2], and ground communications were used to warn of impending gas attacks. The widespread use of radio communications, however, was largely limited by the bulkiness of available field equipment [3]. The broader importance of radio transmissions increased during the latter portions of the twentieth century, facilitating applications such as the direction of troop movements and ordnance targeting, and enabling radar as a key defensive systems for early warning and target recognition. Simultaneously, these same signals were being exploited by adversaries able to derive information such as broadcast location, representing perhaps the first instances of what is currently called electronic warfare (EW) [4]. Similarly, forces began deploying systems that either had specially engineered microwave signatures or actively sought to deceive an opposing force through countermeasure broadcasts and the like.

Two examples of the emergence of EW come from World War II, where British scientists deployed electronic countermeasures (ECM) intended to prevent German bombing raids. In one instance, British scientists detected the German Knickebein navigation beam [2] and in response retransmitted distorted signals, leading to German bombing missions targeting empty fields instead of densely populated urban areas. Separately, in 1943, the British deployed chaff—thin strips of metal foil with length equivalent to half the radar wavelength—which saturated German tracking radars. The deployment of such countermeasures did not end the cycle of escalation, however, as counter–countermeasures soon appeared, leading Winston Churchill to eventually dub these aspects of the war as part of a "Wizard War [2]."

While progress continued in the use of radar and radar jamming techniques during the second half of the twentieth century, the use of electronics outside of this more traditional RF application set exploded with the invention of the transistor, followed by the digital integrated circuit (IC), the monolithic microwave integrated circuit (MMIC), and the microprocessor. Modern defense communications systems rely heavily on point-to-point and satellite radio/microwave links based on IC technology, and routing over complex networks via microprocessors. Beyond communications, electronics are exploited for target identification and tracking, navigation, systems control, and logistics management. Some examples include the development of modern, microprocessor-based, fly-by-wire aviation systems in the 1970s, and "smart-bomb" radar-guided munitions using MMIC technology in the 1980s and 1990s.

12.2 The Classification of Electronics in Military Systems

The ubiquity of electronics in modern defense systems requires the development of a system of categorization for the wide variety of applications utilizing some form of electronic functionality. In this section we provide several operational definitions germane to the topic, a description of the manner in which the military typically categorizes defense electronics applications, and a number of commonly encountered acronyms.

Electronic warfare is broadly defined as the application of electromagnetic spectrum exploitation to military needs, and is derived from a more focused definition that describes EW as the use of an adversary's RF emissions for one's own tactical advantage [5,6]. This is accomplished through a series of actions extending across all phases of battle, from sensing and analysis of an adversary's spectrum use, to the deployment of appropriate ECM and electronic counter–countermeasures (ECCM) as appropriate [4]. As such, EW is designed to benefit the host force by providing spectral dominance at the expense of the adversary's spectrum access.

EW functionalities are subdivided into the following classes: Electronic support (ES), Electronic attack (EA), and Electronic self-protection (EP) [4]. In the following paragraphs a brief overview of these categories is provided, along with examples of systems or functionalities falling into these categories. For a detailed accounting of the organizational structure of EW operational objectives see Reference 7.

The first category, ES, describes the actions undertaken to sense and detect spectral utilization by other parties. ES includes electronic support measures (ESMs), from radar warning receivers (RWRs)—simple, highly reliable devices designed to detect the imminent threat posed by the presence of a hostile radar system tracking friendly platforms—to highly complex ESM platforms designed to map out the total traffic of pulse and CW signals across a broad spectral range [7]. ESM-COM systems intercept enemy communications and locate transmission points. In addition, ES typically encompasses all aspects of signal intelligence (SIGINT), including electronic intelligence (ELINT), communications intelligence (COMINT), and radiation intelligence (RADINT), although these are sometimes categorized independently [8].

EA, otherwise known as ECM, describes actions taken to neutralize an adversary's exploitation of the electromagnetic spectrum by deliberate jamming or deceptive practices [6]. ECM includes active and

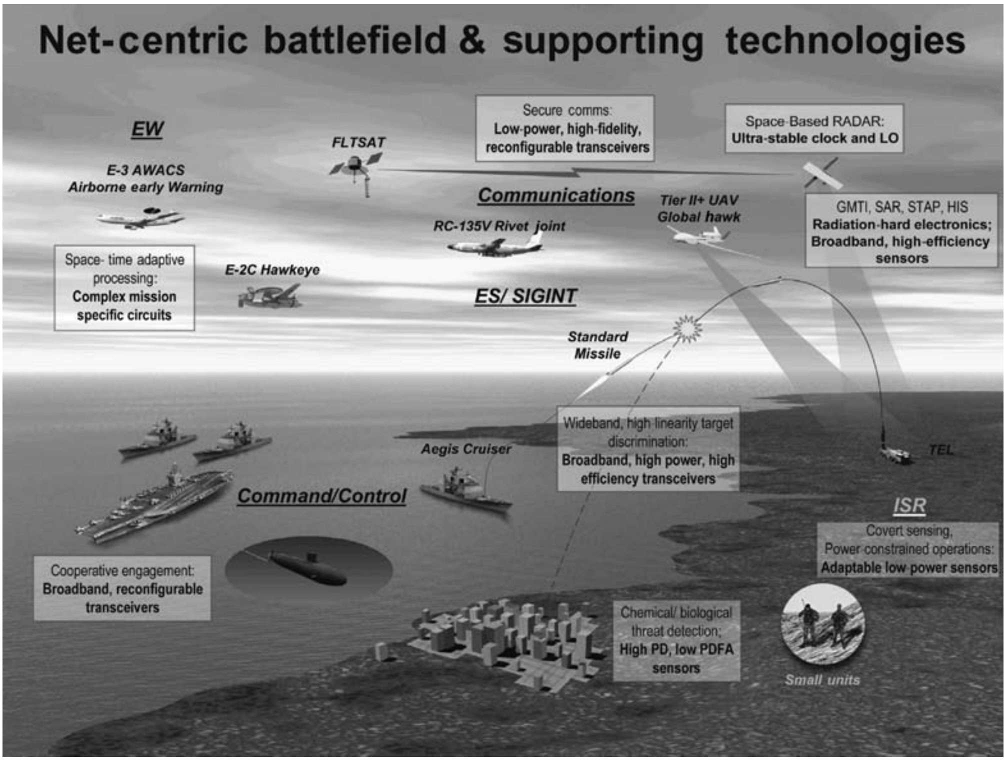


FIGURE 12.1 The net-centric battlefield identifying major platforms, capabilities, and enabling technologies.

passive, onboard and offboard systems, such as the use of chaff to counter radar tracking, stealth to reduce radar signatures, noise or deception jammers to interfere with enemy tracking, and decoy deployment to mislead the targeting party [7]. Finally, EP, also known as ECCMs, are "actions taken to reduce the effectiveness of enemy used ECM to enhance the use of the electromagnetic spectrum for friendly forces [6]." ECCM are deployed in order to allow for the use of standard EW platforms in a hostile environment. Detailed descriptions of ECM and ECCM systems may be found in References 7 and 9.

Although EW is typically defined as it relates to exploitation of the electromagnetic spectrum, it clearly relies upon a broad range of technologies in contemporary analog and digital electronic devices and systems. In fact, it is useful to introduce an alternate organizational perspective which focuses on battlefield operations and "net-centric warfare" instead of explicit spectral dominance (Figure 12.1). This is known as C4ISR [10], or Command, Control, Communications, Computers, Intelligence, Surveillance, and Reconnaissance. It is obvious from the names of these categories that they overlap strongly with many of the subdivisions of EW engagements. However, this heading more accurately reflects the breadth of electronics applications within military systems and operations. For example, the "Control" and "Computers" categories rely heavily on digital microprocessors, although electronics applications form only a small subset of general C4ISR frameworks.

12.3 DoD RF/Microwave Technology Requirements

Military electronics typically operate near the performance envelope, requiring higher power and bandwidth than those used in commercial applications while at the same time requiring superior reliability with precise operational specifications. Although common military systems derive a considerable fraction of their electronic components from commercial-off-the-shelf (COTS) sources, the realization of extreme performance along key system metrics (e.g., range, sensitivity, data rate, and so on) is typically enabled by specialized components. Such components are often highly customized, and due to the nature of the military market, produced in low volume.

12.3.1 Mitigation of Commercial Encroachment

The high power and bandwidth requirements mentioned above result both from extreme operational requirements (such as harsh environment functionality) and the encroachment of commercial systems into the operational phase-space formerly occupied only by the military. Such encroachment is most evident in the utilization of radio bandwidth as will be discussed later, and caused concern for the U.S. Government as early as 1917. At that time, the United States had just entered into World War I, and in order to maintain spectral access for military needs, the U.S. Government ordered the shutdown of all private U.S. radio stations [11]. While the government does not take such extreme measures at present, spectral access remains highly regulated. In cases where spectral bands are shared between commercial and military applications, the military is typically forced to operate at higher power in order to ensure functionality, as is the case for satellite communications, radar, and EW/ECM. Further, increased demand for spectral access has led to tightening frequency tolerances on military components due to diminished channel spacing.

12.3.2 Component Reliability—Radiation Hardness

In addition to spectral access, the DoD is concerned with the reliability of electronic devices for deployment in military systems. This has translated into the Military Specifications or "Mil-Spec" standards for electronic components, but has also led to the development of entirely new technologies designed to address pressing DoD needs. One area of particular concern is radiation hardness; digital and analog ICs are susceptible to degradation due to exposure to radiation, and this degradation must be mitigated. Accordingly, aerospace platforms that are typically exposed to cosmic rays (or may be subject to nuclear fallout), and electronic controls in nuclear energy applications subject to radiation such as neutron flux, require specialized electronic devices and circuits robust under such conditions.

Essentially all of the materials in an IC are in some way radiation sensitive—from the semiconductor materials and gate dielectrics to metal interconnects. The effects of radiation on IC technologies can be described in terms of three main damage categories. First, ionization damage produces electron-hole pairs in the gate dielectric either through the direct absorption of photons, the traversal of charged particles through the circuit, or the action of secondary particles created by incident heavy particles. Although these photoelectrons are typically free to migrate, holes become trapped and significantly degrade transistor performance, increasing leakage current, shifting threshold voltages, degrading gain, and reducing transconductance. When the total ionizing dose reaches a device-specific threshold, that device becomes permanently inoperable. Next, incident particles such as hadrons may produce crystal dislocations by physically displacing semiconductor host or donor atoms. These dislocations serve as scattering sites for charge carriers, thereby increasing device resistance. Finally, single event effects are linked to a circuit being struck by a single, high-energy, ionizing particle. This particle creates an ionization path through the circuit that acts as a temporary short between components. These processes may result in single-event upsets (SEUs) that flip a logic bit value, single-event latchups (SELs), or single-event burnouts (SEBs). For details on the various kinds of radiation damage common to semiconductor devices see Reference 12.

Radiation hardness is required for both digital CMOS and analog or mixed-signal compound-semiconductor technologies, as both appear in aerospace platforms such as satellite communication systems. A number of approaches have been developed for the creation of radiation-hard devices and circuits, including the development of CMOS [13] and GaAs [14] fabrication processes that incorporate additional materials or circuits to create a "radiation-hard-by-process" capability. In CMOS these processes typically rely on the following: the use of an SOI structure with a thin buried oxide that allows for low-mobility holes to tunnel out of the active region of the device, the incorporation of feedback resistors to reduce latch-up susceptibility, or the implementation of redundancy in the circuit topology to reduce the opportunity for single point failures. In addition, new transistor designs and layouts that minimize edge leakage have been developed and deployed. Radiation-hard circuitry also typically incorporates alternate circuit designs, redundancy, and error detection and correction functionality in order to mitigate the effects of SEUs. This last strategy is referred to "radiation-hard-by-design" since it relies only on design changes without altering the standard CMOS fabrication process. In practice, a combination of radiation-hard by-process and by-design techniques are employed to meet DoD requirements. For details and an expanded discussion see References 15 and 16, or find technical material relating to radiation-hard electronics in the context of high-energy physics experiments through [17].

12.4 RF and Microwave Technology in DoD Systems

The previous sections of this chapter provided an organizational framework for DoD electronics applications, as well as details of the requirements placed on electronics for DoD systems. This section will focus on introducing specific electronic technologies, largely at the circuit and sub-system level, that have enabled such systems, and will describe the trends for emerging technologies. Throughout this section we will refer to components and functionalities in the context of ubiquitous military radar systems, but similar analyses are applicable to other RF and microwave systems.

12.4.1 RF and Microwave System Architectures

Historically, many advances in electronic components has been driven by military needs only to be later adapted for commercial applications. For example, many advances in RF technology occurred during World War II with the development of RADAR (radio detection and ranging) systems for early warning of incoming aircraft. The first RADAR systems were based on relatively crude (by today's standards) microwave sources and point detectors. Subsequent attempts to improve RADAR performance led to the discovery and development of new electronics components, including the realization of an improved RF source, the cavity magnetron. Progress continued by increasing the frequency and power ranges accessible using vacuum tube sources. In parallel, improvement in the sensitivity of detecting receivers led to advances in rectifier technology and some of the first semiconductor diodes [18,19]. Here, we describe the advances in RF system architectures associated with progress in radar technology.

Early radar consisted of a central transmitter and receiver architecture with passive beam steering or mechanical scanning of a parabolic dish. The overall performance was determined by the performance of individual components as well as the effects of any losses incurred in routing the signals to and from the radiating aperture. For many years, the radiation was focused and collected solely by a parabolic dish, but it was later realized that moving to a distributed radar architecture based on a phased-array (Figure 12.2a) enabled several system advantages. Foremost among these was the ability to electrically steer the emitted beam through the application of a graded phase shift to each element of the array face in such a way that the far-field beam is determined by the phase steering (Figure 12.3). Such behavior, enabled through the development of the phased-array antenna, was facilitated by the invention of the of phase shifter, an analog component that introduces a determined shift of phase along the signal path.

A further advance, enabled by the development of semiconductor low-noise amplifiers (LNA), was to replace a single common receiver with an array of distributed LNAs at the array face (LNAs are covered in more detail in the next section). Distributing the LNAs at each antenna element improves overall system performance by exploiting array gain and the statistical, uncorrelated nature of background noise. Use of a distributed receiver architecture in the presence of uncorrelated noise sources reduces the noise floor, and

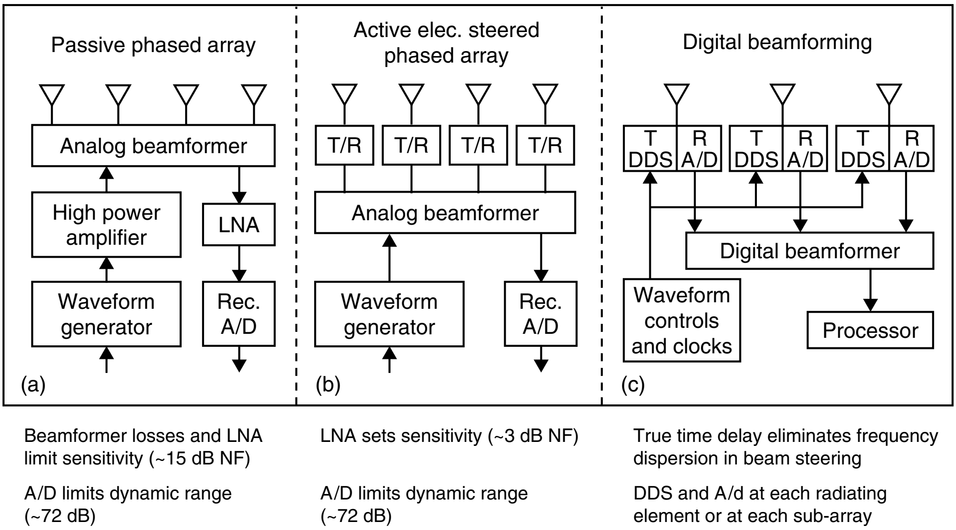


FIGURE 12.2 Primary RF/microwave architectures germane to military applications.

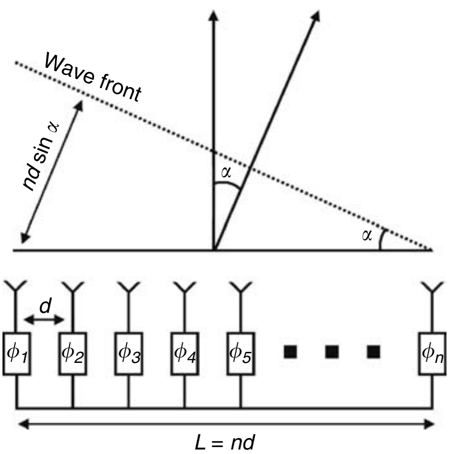


FIGURE 12.3 Schematic diagram of a phase-steered wavefront identifying individual antenna elements with independently controlled phases. The angle of the emitted wavefront, α, is determined by the total phase shift across the array.

hence improves system signal-to-noise ratio (SNR) by a factor proportional to the inverse square root of the number of elements. For example, given a physically reasonable 10,000 element array, the noise floor is reduced by 100 times, an improvement very difficult to achieve through the improvement of individual component performance alone.

Subsequent developments in RF system architecture have involved extending the distributed-element concept to the transmit function as well as the receive operation, as represented by Figure 12.2b. The development of semiconductor high-power amplifiers (HPAs), discussed in detail in the next section, enabled a distributed transmit and receive (T/R) architecture that benefits from array gain during both functions. On the transmit side, the effective radiated power (ERP) delivered by an active aperture array is the product of the individual element power (P_E), the number of elements (N), and the array gain which is also proportional to N

$$\mathrm{ERP} \cong P_E N^2 \tag{12.1}$$

Equation 12.1 is approximate as it does not account for the efficiency of the radiation element and does not include a correction for scanning efficiency [7]. The distributed T/R architectures, commonly referred to as active aperture electrically scanned arrays (AESAs), dramatically improved the performance of RF systems.

Further improvements on the AESA architecture are afforded by exploiting digital processing capability at, or close to, each radiating element. For example, since all backend target identification is accomplished through digital signal processing (DSP), any improvements in the bandwidth or resolution of the digital representation of the analog signal can be exploited at the DSP. The associated need for high-performance mixed-signal components [analog-to-digital converters (ADCs or A/Ds) and digital-to-analog converters (DACs or D/As)] is often the rate-limiting factor for RF-system performance enhancements. However, one approach to relaxing the requirements on the ADC or DAC is to again exploit a distributed array architecture. As was done for the LNA and HPA, if a single ADC at a central processing point is replaced with ADCs located behind each element of an array, the effective dynamic range of the system is increased by the array gain. This is again accomplished by driving down the overall system NF, and assumes uncorrelated noise sources. The same approach can be applied to the transmit operation by placing DACs or direct digital synthesizers (DDS) at each array element to *digitally* generate the transmitted signal, as represented in Figure 12.2c. The digital array architecture also enables the use of time-delay beam steering in place of phase steering, an approach that eliminates the frequency dispersion associated with phase steering of the wave front.

The discussion above has outlined the progression in RF architectures, using radar as an example system, although the same trends are evident for communications and EW systems. The final system architecture is determined based on required performance, platform architectural constraints, and affordability. The following sections will give further details on trends in electronic component technology following trends in the RF-system architectures. The discussion will be separated into the areas of analog-RF and mixed-signal circuits.

12.4.2 Analog RF Electronic Components

A thorough discussion of analog RF electronics must address many components including phase shifters, local oscillators, isolators, filters, and switches, but the primary active components are the LNA and the HPA. This section will focus on the key attributes and technological status of these devices.

12.4.2.1 Low-Noise Amplifiers

The LNA is used to detect and amplify incident RF signals over a defined frequency band. An LNA is primarily characterized by its NF, linearity, and dc power consumption. The NF relates to the additive RF noise imposed on the incident signal in the amplification process, and as such the LNA sets the NF for the entire receive chain. The key requirement for realizing a low-NF LNA is a transistor technology with very low on-resistance and high gain at the operating frequency. One formulation for the minimum NF (NF_{min}) for an LNA based on a field-effect transistor is given by Fukui [20] as

$$NF_{min} = 1 + K_f \frac{f}{f_T}\sqrt{g_m(R_s + R_g)} \tag{12.2}$$

where R_s and R_g are the transistor parasitic source and gate resistances, g_m the transistor transconductance, f_T the unity current gain frequency, f is the operating frequency, and K_f is a characteristic fitting constant. Equation 12.2 shows the importance of minimizing the transistor parasitic resistances and maximizing gain (realizing high f_T) to achieve a low NF. To meet these requirements, the semiconductor material in which the transistors are realized must have high mobility, high channel carrier density, and high saturated electron velocity. Table 12.1 shows the key material parameters for the primary semiconductors used in modern electronics technology. InAs and InSb components may be employed for low-NF applications due to the high carrier mobilities enabling low on-resistance, and high saturated electron velocity

TABLE 12.1 Material Characteristics for Semiconductors Widely Used in the Electronics Industry

	Si (—)	InAs (InSb)	GaAs (AlGaAs/InGaAs)	InP (InAlAs/InGAs)	4H SiC (—)	GaN (AlGaN/GaN)
Bandgap (eV)	1.1	0.36 (0.17)	1.42	1.35	3.26	3.49
Electron mobility (cm^2/Vs)	1.5×10^3	3.3×10^5 (8×10^5)	8.5×10^3 (1.0×10^4)	5.4×10^3 (1.0×10^4)	7.0×10^2	9.0×10^2 ($>2.0 \times 10^3$)
Saturated (peak) electron velocity ($\times 10^7$ cm/s)	1.0 (1.0)	3.0 (5.0)	1.0 (2.1)	1.0 (2.3)	2.0 (2.0)	1.5 (2.7)
2DEG sheet electron density (cm^{-2})	NA	5.6×10^{12} (1.0×10^{12})	$<4.0 \times 10^{12}$	$<4.0 \times 10^{12}$	NA	$1.0–2.0 \times 10^{13}$
Critical breakdown field (MV/cm)	0.3	0.04 (0.001)	0.4	0.5	2.0	3.3
Thermal conductivity (W/cm-K)	1.5	0.27 (0.18)	0.5	0.7	4.5 (3.3)[a]	>1.7 [21]
Relative dielectric constant	11.8	14.6 (17.7)	12.8	12.5	10	9.0

[a] Thermal conductivity of Semi-insulating SiC.

enabling large f_T. However, as will be described next, NF is not the only important performance metric for LNA.

When receiving or transmitting an RF signal, it is important that the amplification process does not generate additional signals as a result of nonlinear interactions. Linearity is the attribute of an LNA (or HPA) that characterizes the level of amplifier signal input which yields a predefined level of nonlinear interaction, either through harmonic mixing or nonlinear, amplitude-dependent amplification. Specific definitions for linearity are developed for distinct modulation formats such as CDMA or EDGE as used in commercial cellular systems, but the most general definition is based on the third-order intermodulation (TOI) product or third-order intercept point (IP3)—the theoretical point where the desired signal and the third-order distortion have equal magnitudes [22]. Higher TOI allows an amplifier to receive or transmit a signal with relatively lower distortion. Often, linearity can be improved at the expense of dc power dissipation (P_{DC}), so an appropriate figure-of-merit for an LNA in particular is TOI/P_{DC}. The transistor linearity may be optimized through proper design of the transport channel and by increasing the breakdown voltage.

12.4.2.2 High-Power Amplifiers

In an RF system, dynamic range and sensitivity are dominated by the previously described receiver NF and linearity, and the high-power transmit amplifiers. Semiconductor amplifiers have become the dominant HPA technology in active aperture arrays and cellular phones. Military radar systems have some of the most stringent HPA performance requirements due to the need for high power, bandwidth, and efficiency. For example, an X-band radar system with a center frequency at 10 GHz might have a bandwidth well over 1 GHz, whereas a commercial cellular base station would typically have a center frequency near 2 GHz and a bandwidth on the order of 100 MHz or less. The higher operating frequency of the X-band radar stresses the frequency performance of the HPA transistor, while the transistor is simultaneously required to deliver high power. Fundamentally, the available power (P_{out}) from a semiconductor HPA operated in class A is given by

$$P_{out} = \frac{I_{max} V_{max}}{8} \tag{12.3}$$

where I_{max} and V_{max} are the maximum current and voltage of the output transistor, respectively. Although some modification of transistor design can lead to enhancement in I_{max} and V_{max}, the intrinsic power-handling capability of the transistor is determined by the selection of the semiconductor material itself; I_{max} is determined by the carrier mobility and sheet charge density, while V_{max} is defined by the material's intrinsic critical electric field.

For military radar systems operating below 1 GHz, the first semiconductor amplifiers used Silicon transistors. In the 1980s and 1990s gallium arsenide (GaAs) materials were developed that provided operational capability at higher frequencies, and today most high-performance X-band radars for fighter aircraft, Naval warships, and ground-based platforms use GaAs HPAs in active aperture arrays. This technology was developed by the U.S. Department of Defense for such applications but subsequently found wide application in high efficiency HPAs for cellular phones. Although the basic materials are often the same, the transistors and circuits are optimized differently between defense and commercial applications. In addition, military applications requiring high-frequency operation have incorporated indium phosphide (InP) based amplifiers that exploit the high cut-off frequency of InP transistors.

To advance the performance of military RF systems, research is underway to develop HPAs based on a new class of materials referred to as wide-bandgap semiconductors. The bandgap, or energy gap (E_G), is a characteristic property of a semiconductor that determines the critical electric field the semiconductor can withstand. Generally, wide-bandgap semiconductors are defined as having $E_G > 2.5$ eV.

Both silicon carbide (SiC: $E_G = 3.26$ eV) and gallium nitride (GaN: $E_G = 3.4$ eV) have been developed for RF devices. GaN has a critical electric field 10 times higher than Silicon, and seven to eight times higher than GaAs or InP. The higher critical electric field, at which carriers are generated in the material by avalanche multiplication, enables a 5- to 10-fold increase in operating voltage, enabling higher power-handling capability and efficiency. In addition, GaN transistors exhibit a higher sheet current density than GaAs or InP, further increasing their power-handling capability. Current research results have demonstrated unit power densities for GaN that are up to 20 times higher than GaAs [23]. This technology is also being developed for commercial wireless base stations, as its high linearity may enable increased spectral efficiency of cellular networks [24]. The remaining key technical challenge for these materials is improving the core HPA reliability as required for military and commercial applications.

12.4.3 Mixed-Signal Electronics

The move to AESAs and digital waveform control has increased the performance requirements on mixed-signal circuits. Military applications stress bandwidth, resolution, and power capabilities across a system, while the most stringent performance requirements are placed on the A/D converter and DDS.

12.4.3.1 Analog-to-Digital Converters

An ADC receives as input an analog waveform and produces a digital representation of that signal as output. An excellent review of ADC performance and limitations was given by Walden in Reference 25. The performance of an ADC is typically characterized by the following parameters: sampling rate, resolution, and power. Sampling rate is typically set by the Nyquist condition whereby data are sampled at twice the signal bandwidth to ensure capture of all signal information ($f_{signal} = f_{sample}/2$). The signal bandwidth, f_{signal}, also known as the effective-resolution bandwidth (ERBW), is defined as the frequency at which the signal amplitude decreases by 3 dB relative to the low-frequency amplitude.

The resolution of an ADC is limited by noise phenomena that can be grouped into four categories: quantization noise, thermal noise, comparator ambiguity, and aperture uncertainty. Quantization noise refers to the error introduced in the discretization of a continuously varying analog signal. The thermal noise is the spectral noise density seen at the ADC input due to Johnson–Nyquist voltage fluctuations. Comparator ambiguity results from the finite time required by a transistor in the comparator to respond to small voltage differences in determining if the input signal has exceeded the comparator reference voltage. Accordingly, one motivation for increasing transistor switching speed is to reduce comparator ambiguity. Finally, aperture uncertainty arises from jitter in the ADC's sampling interval. This timing jitter introduces an error in the sample that varies with sampling frequency. In practice, the sampling interval is represented by a mean value and standard deviation.

The input frequency and the ADC resolution determine the maximum tolerable aperture jitter (τ_a) as

$$\tau_a = \frac{1}{2^N \pi f_{max}} \tag{12.4}$$

where N is the stated number of bits of resolution (different from the effective number of bits, ENOB, which account for noise and distortion effects) and f_{max} is the maximum input frequency [26].

ADCs have also become important for commercial wireless and fiber communications, but commercial applications tend to have more lenient resolution requirements relative to military applications. A nice review of recent progress is given by Le et al. [26]. In addition, the paper of Merkel and Wilson gives a review of ADCs in the context of unique military applications [27].

12.4.3.2 Direct Digital Synthesizers

As discussed above, a recent trend in RF system architecture is toward the use of high-frequency DDS for digital synthesis of the RF output waveform. Recent progress in high-speed InP heterojunction bipolar transistor (HBT) performance demonstrating transistor switching speeds over 450 GHz, is making possible the realization of DDS with output in the microwave bands above 10 GHz [28]. The usefulness of a DDS is determined by its phase noise characteristics, as well as its spur free dynamic range (SFDR), which measures the ability to generate a clean central frequency without undesirable intermodulation products. Increasing SFDR is a critical challenge that is being addressed through improvement of the transistor dynamic range and linearity, as well as the development of improved circuit architectures for high-speed DDSs.

12.4.4 Emerging Technology Trends: Reconfigurable Electronics and RF Photonics

As RF military systems continue to demand improved performance and increased agility in functionality for platforms restricted in size and power, there has emerged a growing trend to develop reconfigurable and adaptable RF components. The performance of reconfigurable components can be changed to a preset combination of distinct operating points via external control. Examples of this design paradigm are reconfigurable RF amplifiers with dynamically tunable on-chip input and output matching networks that can be optimized over a wide range of frequencies [28]. Such functionality eliminates the normal trade-off a circuit designer must make between power, efficiency, or gain and operating frequency. It also allows the system engineer to further optimize *system* performance after hardware assembly by adjusting the operating points of the various subcomponents. This may be important for overcoming undesirable parasitic effects at the module level, and for accommodating changes in antenna impedance-matching conditions that result from varying signal environments.

Beyond reconfigurable components are adaptable components that possess some degree of autonomy to self-optimize, test, or monitor without external intervention. This requires embedded control and sensing derived from the co-optimization of hardware and control algorithms. Such capabilities are already being incorporated into complex digital circuits, as an increasing number of transistors are being used to monitor central operations, perform error correction, and potentially modify the operation of circuit blocks to optimize performance [28].

Finally, RF photonics refers to the propagation of signals at microwave frequencies over optical fiber by imposing an analog RF signal on an optical carrier [29,30]. This technique, understood for many years, has the attractive feature of allowing one to remote the radiating antenna element from some or all of the high-performance electronics. Further, the inherent high speeds allowed by photonic transport at telecom frequencies allow extremely broad bandwidth signals to be transmitted. The challenges of employing RF photonics have been noise introduced on the RF signal and nonlinear signal generation that limits optical link linearity and dynamic range. The RF noise has been primarily attributed to losses introduced through the modulation of optical carriers using a Mach Zehnder optical modulator, but noise is also introduced in the optical detectors and in optical generation at the reference carrier laser. All of these noise components are being addressed in ongoing research efforts, and RF photonic links are now close to realizing net gain. When this is achieved, remoting of RF apertures will become a viable design paradigm for future systems. For an overview of recent progress in RF photonic technology see Reference 31.

12.5 DoD Applications

Using knowledge of the organizational frameworks, military requirements, and technologies described above, this section addresses a few of the key defense applications for RF and microwave electronics technologies. This list is not meant to be exhaustive, but instead illustrative of some of the major thrusts within the DoD. Detailed explanations of background material and underlying premises are left to other sections of this handbook.

12.5.1 Military Use and Allocation of the EM Spectrum

During the development of radar and its deployment for defensive and offensive purposes during World War II, there was only limited use of the radio spectrum, and accordingly bandwidth and frequency were readily available to the military. With the information revolution and the explosion of inexpensive consumer electronics, DoD has faced significant challenges associated with the encroachment of commercial users into traditionally government-only spectral bands. At the same time, military spectrum use is expected to grow 70% in the first decade of the twenty-first century, and military data transmission requirements are increasing exponentially with time [32]. In terms of hardware, the DoD maintains an expansive inventory of equipment designed to exploit the RF spectrum, totaling over 800,000 active units worth approximately $100B [33]. Accordingly, ensuring access to the EM spectrum for active transmitting and receiving devices is a primary concern for the military.

Military applications such as radar, communications, and navigation rely heavily on the ability to transmit EM signals as will be described in the following sections. Owing to heavy military and commercial use, regulation of access to, and functional organization of, the EM spectrum is required. The allocation of spectral bands in the United States is handled by both the Federal Communications Commission (FCC) and the National Telecommunications and Information Administration Office of Spectrum Management [34], while the International Telecommunications Union (ITU) [35] assigns frequency bands specifically for radar use [9].

The Institute for Electrical and Electronic Engineers (IEEE) and NATO have both created classification systems to divide up the RF and microwave spectrum into functional categories. The IEEE system is based on the code names for frequency bands introduced during World War II, while the EU/NATO have adopted a more modern convention for binning over the same spectral range. In Figure 12.4, we present both the general classification of radio bands from extreme low frequency (ELF) through extreme high-frequency (EHF), and point out some of the most important military uses for these bands [7]. In addition, we provide the IEEE and NATO RF band designations, as well as ITU radar bands.

12.5.2 Active Transmitting Systems

This subsection introduces some of the main application areas in which the military exploits access to the EM spectrum. Here, we provide an overview of radar, communications, and navigation techniques of interest to the military.

12.5.2.1 Radar

The use of electromagnetic radiation to locate, track, and identify objects is of obvious utility to defense applications, and first began to play a major role in warfare during World War II. In a system with such

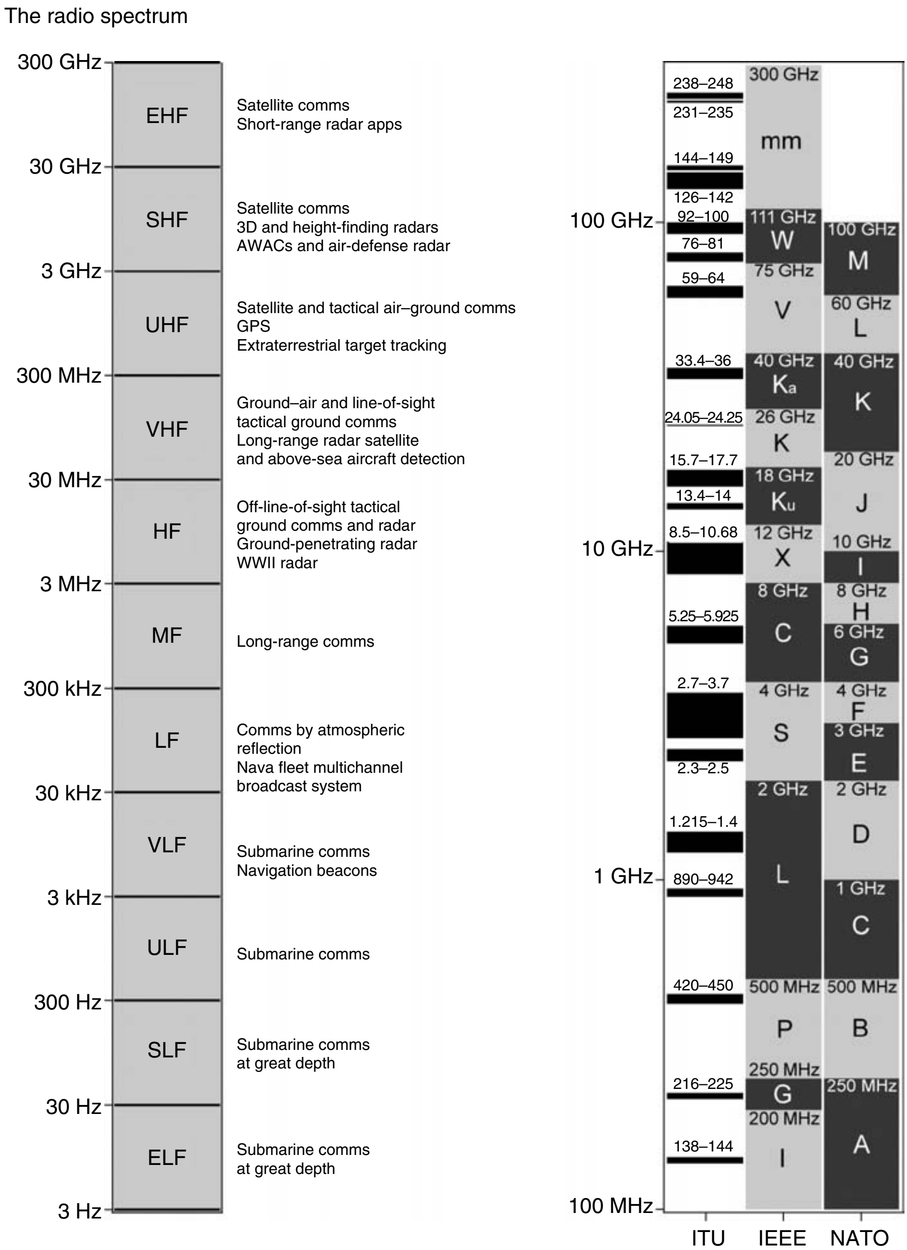


FIGURE 12.4 Military use of the radio spectrum and official designations of frequency bands important for military applications. White sections indicated unallocated frequencies, while black and grey are both used to denote assigned bands for clarity.

functionality, radiation is generated and emitted through a directional antenna, and reflected energy is detected. Simultaneously, the time-delay between emission and detection is recorded, providing a measure of target distance. Further, Doppler shifts in the reflected signal are used to distinguish moving targets from a stationary background, or even moving components of a single target, given sufficient bandwidth (for high-range resolution) and antenna size (for angular resolution) [9]. In fact, Doppler shifts may be exploited to provide enhanced angular resolution without the need for extremely (electrically) large antennas in synthetic aperture radar (SAR) systems.

In addition to standard tracking radar systems familiar from air traffic control in civilian use, many weapons systems employ radar for tracking or target identification. For example, on the battlefield, targets are identified by a central warning center, and these threats are assigned to fire control centers responsible for target tracking and neutralization via simple artillery or more complex missile systems. Short-range command missiles are actively tracked and guided by the target center, and as such do not interact with the target themselves [7]. The invention of MMICs allowed for the mounting of light-weight, low-cost radar systems (active and passive) on conventional munitions in order to provide guidance for precision targeting, and in order to move enhanced capabilities from the fire center to the munitions themselves. Systems with such functionality include beam-riding missiles that follow a tracking beam to a target using an onboard sensor, passive antiradiation missiles that lock onto radar sources, semi-active homing missiles that track a reflected radar signal from a target, but do not contain their own transmitters, and active homing missiles that utilize an internal active guidance and tracking system [7].

Of course, many other military capabilities are enabled by radar functionality, such as the detection of underground structures, collision avoidance for aircraft or land vehicles, and meteorology. For a comprehensive overview of radar systems and applications, see References 7 and 9 and other sections of this handbook.

12.5.2.2 Communications

As discussed in the introduction to this chapter, RF and microwave propagation through free space is used as a means of communications between distant parties in the military's C3I (command, control, communications, intelligence) framework. The ability to encode messages on RF and microwave transmissions has been well known for decades and exploited extensively in military campaigns. Single stage radio and microwave links are typically used for short-range ($\sim$40 miles), point-to-point communications such as those in two-way radios, air–ground, and air–sea links. These channels may employ line-of-sight propagation paths or indirect paths requiring scatter or diffraction from obstacles, although these typically lead to high loss. Long-range communications typically rely on microwave relay systems where signals are received, amplified, and retransmitted. Alternatively, it is possible to use scattering of microwaves from the troposphere or ionosphere in order to implement non-line-of-sight links over long distances (several hundred miles), although as in diffractive or scattering assisted point-to-point radio links, losses are high and special techniques are typically required. Beyond these techniques, one may employ satellite-based communications systems whereby two distant points separated by up to 8000 miles are indirectly connected via an orbiting satellite which receives, amplifies, and transmits uplinks in the microwave region [22]. It is worth noting that modern military systems rely on encrypted digital signals and spread spectrum techniques [36] for the establishment of secure and robust communications channels.

12.5.2.3 Navigation

Satellite systems have also become increasingly important for navigation. The most common example is the global positioning system (GPS) [37], whose receiver units are deployed throughout the military, even at the level of the individual soldier. All active radio navigation systems are based on a simple sequence: local unit broadcasts precisely timed pulses at frequency f_1, fixed station with known position receives and retransmits at frequency f_2, local unit receives and extracts distance to fixed station based on the travel time of transmitted and received pulses [22]. Triangulation between three base stations allows precise determination of altitude, latitude, and longitude. Microwave carrier signals and space-borne platforms replaced older very low frequency (VLF) ground-based systems such as Omega [22], as the former provides enhanced accuracy and resolution with an unobstructed view of the earth's surface. Military systems are capable of determining user position with accuracy better than 10 feet.

12.5.3 Sensor Systems

This section describes the role of sensors in detecting electromagnetic radiation in military systems. A discussion of RF and microwave sensors as applied to EW in ESM systems is provided, and is followed by examples of specific sensing systems and modalities of importance to the military.

12.5.3.1 ESM Sensors

Passive sensor systems capable of detecting hostile EM spectrum exploitation typically fall under the designation of ESM. ESM sensor systems are designed to provide surveillance capabilities to general EW architectures by searching for and receiving EM signals, and locating the sources of EM radiation. These systems must be able to intercept and analyze a variety of signals with large bandwidth and dynamic range in order to provide useful information with short time lag, over a wide spectral range, and in the presence of clutter [4]. Traditional ESM systems typically operate from the HF through 40 GHz [8] for the detection of enemy radar, or from the ELF through the UHF for the interception of communications (ESM-COM) [7] (see Figure 12.4). Further, these systems generally possess no knowledge of a given threat signal a priori, and therefore must be able to analyze across multiple parameters (carrier frequency, power, dynamic range) in real time. For example, radar interception requires 1 MHz frequency resolution with coverage from 0.5 to 18 GHz, sensitivity better than –60 dBm, and dynamic range >50 dB [8].

One may derive the received power in an ESM sensor from the Friis Transmission equation as

$$P_R = G_R(P_T G_T)\left(\frac{\lambda_0}{4\pi R}\right)^2 \tag{12.5}$$

with G_R the receiver gain, P_T the transmitter power, G_T the transmitter gain, λ_0 the radiation wavelength, and range R [22]. It is thus straightforward to extract the range of such a detector by simply setting a minimum received power and solving for R. ESM sensor implementation relies upon signal collection by both omnidirectional and direction-finding antennas, and the use of MMIC technology for high-bandwidth receivers [7,8]. Common receiver technologies employed in ESM sensors include wide-open frequency, superheterodyne, channelized, microscan (compressive), crystal-video, acousto-optic, and digital receivers. The actual functionality of a given receiver system will be based on the chosen application and can vary significantly. For a complete discussion of ESM sensor systems and their related component technologies see References 7 and 9.

12.5.3.2 Sensor Networks

Sensors for ESM applications typically do not act independently, but are instead integrated in a network in order to aid in source location and early warning [4]. Such networks may be housed on a single platform such as a ship or aircraft, or may be connected by a wide-area network spanning the battlefield (see Figure 12.1). Sensors are connected by telecommunication links and positioned such that their fields of coverage overlap, and as with navigation systems, triangulation between multiple sensors allows for the location of a signal source. Further, a single airborne platform can be used for rapid triangulation of an emitter by measuring an enemy signal from multiple locations [7]. At present, considerable effort is being invested in the field of wireless sensor networks. In these networks, power-starved sensors are capable of detecting environmental information and relaying data between network nodes by means of radio broadcasts For a summary of the technological challenges and requirements associated with these systems see References 38–40.

12.5.3.3 Example Sensor Systems

12.5.3.3.1 Radar Warning Receivers

An example of a widely deployed ESM sensor system is the radar warning system typically installed on aircraft in order to detect, and alert the pilot to the presence of, a locked radar tracking signal. In general, hostile radar waveforms, frequencies, and so on are preprogrammed into such a system [5]. Lock-on is recognized by the persistent presence of a constant signal, while radar in search-mode will appear amplitude modulated as it is swept across the aircraft. The standard requirements of RWRs are that the radar signature can be detected at a range larger than that of the associated weapon system, and that the RWR is highly reliable with low false-alarm rate in a cluttered environment. Again, RWRs may use crystal-video, swept-narrowband-superheterodyne, tuned-RF, or wideband-superheterodyne receivers [7]. For performance comparisons and detailed description of these technologies see References 5, 7, and 41.

12.5.3.3.2 Radiometic Detectors

Radiometry systems rely on blackbody emission in order to perform temperature mapping of a given location or object, and are widely used in the military for surveillance, mapping, and the monitoring of surface conditions from space-based satellite platforms. A blackbody is defined as absorbing all incident energy (while reflecting none) and emitting an equal amount of energy. Real objects are non-ideal, however, and so we define the radiated power as $P = \varepsilon kTB = kT_B B$, with k Boltzmann's constant, $0 \leq \varepsilon \leq 1$ the object's emissivity, T the body's temperature, $T_B = \varepsilon T$ the object's brightness temperature, and B the bandwidth [22].

A radiometric system typically contains a low-noise receiver which measures radiated power within a narrow band to determine an object's brightness temperature. Because of the constraints on the value of the ε, all non-ideal radiators have brightness temperatures lower than their actual temperatures. Variation in the emissivity, and hence the measured brightness temperature, between objects in a given scene allows for the assembly of a radiometric map, even for objects in thermal equilibrium.

While blackbody radiation at ambient temperatures has a spectral power heavily weighted to the infrared, radiometric systems operating in the microwave, millimeter-wave and submillimeter-wave regions are widely used as they are able to image through smoke or clouds [22]. Using such systems it is routinely possible to achieve temperature resolution of $\sim$1 K.

12.5.4 Other Applications

Electronic devices have grown so common in military systems that to account for all of them would be nearly impossible. Although this chapter has attempted to point out some of the most important military applications exploiting RF and microwave technology, it has not been all-encompassing. A few noteworthy applications are mentioned above, and others illustrated in Figure 12.1, but further discussion is left to other authors. The following references are suggested for interested readers

- Avionics: The "Glass Cockpit [42]" and "Fly-by-wire [43,44]."
- Digital computers and networks for command and control and net-centric warfare/global information grid [45].
- RFID for inventory and logistics management [46–48].
- Ultra-precise time/frequency standards for satellite systems, EW, and so on [49–53].

12.5.5 JETDS

In 1943, a common system of nomenclature and standardization for DoD electronics was developed and adopted by the Joint Communications Board for Joint Army-Navy use [54]. This system, originally designated the Joint Army-Navy Nomenclature System (AN System), was later expanded in scope beyond these two services. The adoption of this system spread between multiple services and even to Canada before the Department of Defense instituted the MIL-STD-196 "Joint Electronics Type Designator System" in 1957. As the system matured it was adopted by the National Security Agency and later Australia, Great Britain, and New Zealand. The JETDS MIL-STD-196 has been redeployed in multiple versions with the latest, MIL-STD-196E being approved by DoD in 1998 [54]. It is used within the U.S. DoD for the classification of RF/microwave, digital, and similar electronic devices and systems. The format of this naming convention is included for completeness of discussion.

The system employs names in the following structure: AN/$X_1X_2X_3$—##X_4. In this format, AN always stands for Army/Navy, representing the historical origins of the system. The next three letters represent the equipment installation (X_1), type (X_2), and purpose (X_3). These letters are followed by an Arabic numeral (##) which corresponds to the particular equipment listing in that category. Finally, in some instances an additional letter (X_4) appears at the end of the tag to denote modifications which maintain

TABLE 12.2 JETDS Naming System and Classification Categories, After Reference 54

MIL-STD-196E AN/$X_1X_2X_3$—##X_4 Installation (X_1)	Type (X_2)	Purpose (X_3)	Miscellaneous (X_4)
A. Piloted Aircraft	A. Invisible Light, Heat Radiation	A. ** Auxilliary assembly	X, Y, Z, Changes in voltage, phase, or frequency
B. Underwater Mobile, submarine	B. *Comsec	B. Bombing	T Training
C. *Cryptographic	C. Carrier–Electronic Wave/Signal	C. Communications (receiving and transmitting)	(C) NSA use only
D. Pilotless Carrier	D. Radiac	D. Direction Finder, Reconnaissance and Surveillance	(P) Units accepting plug-ins
F. Fixed Ground	E. Laser	E. Ejection and/or Release	(V) Variable items
G. General Ground Use	F. Fiber Optics	G. Fire Control or Searchlight Directing	(-FT,-IN) Identical items with varying lengths
K. Amphibious	G. Telegraph or Teletype	H. Recording/Reproducing	Automatic Data Processing (ADP)
M. Mobile (Ground)	I. Interphone and Public Address	K. Computing	1. Digital Equipment only
P. Portable	J. Electromechanical or Inertial Wire Covered	M. Maintenance/Test Assemblies	2. Analog Equipment only
S. Water	K. Telemetering	N. Navigational Aids	3. Hydrid (1 and 2 combined)
T. Transportable (Ground)	L. Countermeasures	Q. Special or Combination	4. Input/Output Device
U. General Utility	M. Meteorological	R. Receiving/Passibe Detecting	5. Magnetic Media
V. Vehicular (Ground)	N. Sound in Air	S. Detecting/Range and Bearing, Search	6. Others
W. Water Surface and Underwater combined	P. Radar	T. Transmitting	Single asterisk (*) are for National Security Agency (NSA) use only
Z. Piloted-Pilotless airborne vehicles combined	Q. Sonar and Underwater Sound	W. Automatic Flight or Remote Control	Double asterisks (**) are for Department Control Point use only
	R. Radio	X. Identification and Recognition	
	S. Special or Combination	Y. Surveillance (search, detect and multiple target tracking) and control (both fire control and air control)	
	T. Telephone (Wire)	Z. *Secure	
	V. Visual and Visible Light		
	W. Armament (peculiar to armament not otherwise covered)		
	X. Facsimile or Television		
	Y. Data Processing or computer		
	Z. *Communications		

interchangeability [9,54]. Major categories for each entry are listed in Table 12.2. For a complete listing of MIL-STD-196E equipment categories and designations see Reference 55.

12.6 Conclusion

The U.S. Department of Defense has always relied on leading edge electronics to deliver unrivaled performance in its systems. RF and microwave systems for communications, radar, remote sensing, EW, and so on are continuously being pressed to deliver enhanced capability at lower power, and in smaller form factors. For this reason, the U.S. DoD is continuing to develop analog-RF and mixed-signal electronics, which typically find applications satisfying needs of both the military and commercial markets.

References

1. Smithsonian Institution. *Cher Ami—World War I Carrier Pigeon.* Available from: http://www.si.edu/resource/faq/nmah/cherami.htm
2. R. Schroer, Electronic warfare. *IEEE Aerospace and Electronic Systems Magazine*, 2003. **18**(7): pp. 49–54.

3. U.S. Signal Corps. *Radiotelegraphy*. Available from: http://earlyradiohistory.us/
4. A.M. Spezio, Electronic warfare systems. *IEEE Transactions on Microwave Theory and Techniques*, 2002. **50**(3): pp. 633–644.
5. D.C. Schleher, *Introduction to Electronic Warfare*. 1986, Norwood, MA: Artech House.
6. R.D. Hayes, Electronic warfare and countermeasures, in *The RF and Microwave Handbook*, M. Golio, Editor. 2001, CRC Press: Boca Raton, pp. 196–215.
7. F. Neri, *Introduction to Electronic Defense Systems*. 2001, Norwood, MA: Artech House.
8. J.H. Collins and P.M. Grant, A review of current and future components for electronic warfare receivers. *IEEE Transactions on Microwave Theory and Techniques*, 1981. **MTT-29**(5): pp. 395–403.
9. M.I. Skolnik, ed. *Radar Handbook*. 1990, New York: McGraw-Hill.
10. c.f. www.c4isrjournal.com
11. T.H. White. *Radio During World War One (1914–1919)*. Available from: http://earlyradiohistory.us
12. M. Dentan. *Radiation Effects on Electronic Components and Circuits*. Available from: http://rd49.web.cern.ch
13. S.T. Liu, W.C. Jenkins, and H.L. Hughes, Total Dose Radiation Hard 0.35 micron SOI CMOS Technology. *IEEE Transactions on Nuclear Science*, 1998. **45**(6): pp. 2442–2449.
14. H. Derewonko, et al. Evaluation of GaAs low noise and power MMIC technologies to neutron, ionizing dose and dose rate effects, in *Proc. of Third European Conference on Radiation and its Effects on Components and Systems*. 1995: IEEE.
15. T.P. Ma and P.V. Dressendorfer, *Ionizing Radiation Effects in MOS Devices and Circuits*. 1989, New York: Wiley-Interscience.
16. G.C. Messenger and M.S. Ash, *The Effects of Radiation on Electronic Systems*. 1993, New York: Van Nostrand Reinhold.
17. *CERN-LHCC RD49 Project*. Available from: http://rd49.web.cern.ch
18. For a description of the development of RADAR and related technology see the following reference.
19. R. Buderi, *The Invention that Changed the World: How a Small Group of Radar Pioneers Won the Second World War and Launched a Technological Revolution*. 1996, New York: Touchstone.
20. H. Fukui, *Low-Noise Microwave Transistors and Amplifiers*. 1982, New York: John Wiley & Sons.
21. V.M. Asnin, et al., High spatial resolution thermal conductivity of lateral epitaxial overgrown GaN/sapphire (0001) using a scanning thermal microscope. *Applied Physics Letters*, 1999. **75**(9): pp. 1240–1242.
22. K. Chang, *RF and Microwave Wireless Systems*. 2000, New York: John Wiley & Sons.
23. Cree Inc. *Cree Demonstrates 400 Watts of RF Power for GaN S-Band Transistor*. Available from: http://www.cree.com/press/press_detail.asp?i=1150225182604
24. Y.F. Wu. 2006, Defense Research Council.
25. R.H. Walden, Analog-to-digital converter survey and analysis. *IEEE Journal on Selected Areas in Communications*, 1999. **17**(4): pp. 539–550.
26. B. Le, et al., Analog-to-digital converters. *IEEE Signal Processing Magazine*, 2005. **22**: pp. 69–77.
27. K.G. Merkel and A.L. Wilson, A survey of high-performance analog-to-digital converters for defense space applications. *Proc. IEEE Aerospace Conference*, 2003. **5**: pp. 5_2415–5_2427.
28. J.C. Zolper and M.J. Biercuk, The path to adaptive microsystems. *Proceedings of SPIE*, 2006. **6232**: p. 1.
29. C. Cox, *Analog Optical Links: Theory and Practice*. 2004, Cambridge, UK: Cambridge University Press.
30. W. Chang, *RF Photonic Technology in Optical Fiber Links*. 2002, Cambridge, UK: Cambridge University Press.
31. C. Cox, et al., Limits on the performance of RF-over-fiber links and their impact on device design. *IEEE Transactions on Microwave Theory and Techniques*, 2006. **54**: pp. 906–920.
32. U.S. Department of Commerce, *Federal Radar Spectrum Requirements*. 2000.
33. Defense Science Board Task Force on DoD Frequency Spectrum Issues, *Coping with Change: Managing RF Spectrum to Meet DoD Needs*. 2000.

34. Department of Commerce National Telecommunications and Information Administration. Available from: http://www.ntia.doc.gov/
35. International Telecommunications Union. Available from: http://www.itu.int
36. P.F. Sass, Why is the army interested in spread spectrum? *IEEE Communications Magazine*, 1983. **21**(4): pp. 23–25.
37. I.A. Getting, The Global Positioning System. *IEEE Spectrum*, 1993. **30**: pp. 36–47.
38. I.F. Akyildiz, et al., Wireless Sensor Networks: A survey. *Computer Networks*, 2002. **38**: pp. 393–422.
39. Christian C. Enz, Nicola Scolari, and Uroschanit Yodprasit. Ultra low-power radio design for wireless sensor networks, in *IEEE International Workshop on Radio-Frequency Integration Technology*. 2005, Singapore: IEEE.
40. M. Aboelaze and F. Aloul. Current and future trends in sensor networks: a survey, in *IFIP International Conference on Wireless and Optical Communications Networks*. 2005, Piscataway, NJ: IEEE.
41. H.J. Belk, J.D. Rhodes, and M.J. Thornton, Radar warning receiver subsystems. *Microwave Journal*, 1984. **27**: p. 199.
42. W. Sweet, The glass cockpit—flight deck automation. *IEEE Spectrum*, 1995. **32**(9): pp. 30–38.
43. R.P.G. Collinson, Fly-by-wire flight control. *Computing and Control Engineering Journal*, 1999. **10**(4): pp. 141–152.
44. S. Osder, Integrated flight/fire control for attack helicopters. *IEEE Aerospace and Electronic Systems Magazine*, 1992. **7**(1): pp. 17–23.
45. J. Morrow and J. Laub. Dynamic policy enablement and management in net-centric warfare systems and communications, in *IEEE Military Communications Conference (MILCOM)*. 2005, Piscataway, NJ: IEEE.
46. J. Landt, The history of RFID. *IEEE Potentials*, 2005. **24**(4): pp. 8–11.
47. US Department of Defense, *DoD Suppliers' Passive RFID Information Guide*. 2005.
48. R. Weinstein, RFID: a technical overview and its application to the enterprise. *IT Professional*, 2005. **7**(3): pp. 27–33.
49. J.R. Vig, Military applications of high accuracy frequency standards and clocks. *IEEE Transactions on Ultrasonics, Ferroelectrics and Frequency Control*, 1993. **40**(5): pp. 522–527.
50. L.L. Lewis, An introduction to frequency standards. *Proc. of the IEEE*, 1991. **79**(7): pp. 927–935.
51. N.D. Bhaskara, et al. On-orbit performance of Milstar rubidium and quartz frequency standards, in *Proc. IEEE International Frequency Control Symposium*. 1997, Piscataway, NJ: IEEE.
52. *Proc. of the 6th Symposium on Frequency Standards and Metrology*, P. Gill, Editor 2002: World Scientific.
53. L. Hollberg, et al., Optical frequency standards and measurements. *IEEE Journal of Quantum Electronics*, 2001. **37**(12): pp. 1502–1513.
54. U.S. Department of Defense, *Joint Electronics Type Designation System*. 1998.
55. Andreas Parsch. *Designations of U.S. Military Electronic and Communications Equipment*. Available from: http://www.designation-systems.net/usmilav/electronics.html

13

Microwave and RF Avionics Applications

James L. Bartlett
Rockwell Collins

Avionics was originally coined by contracting Aviation and Electronics, and this term has gained widespread usage over the years. Pilots of aircraft have three primary necessities; to Aviate (fly the aircraft), to Navigate, and to Communicate. Avionics addresses all three of these needs, but the navigation and communications aspects are where the radio frequency (RF) and Microwave applications lie. Avionics applications typically require functional integrity and reliability that is orders of magnitude more stringent than most commercial wireless applications. The rigor of these requirements is matched or exceeded only by the requirements for space and/or certain military applications. Avionics must function in environments that are more severe than most other wireless applications as well. Extended temperature ranges, high vibration levels, altitude effects, including corona, and high-energy particle upset (known as single event upset) of electronics are all factors that must be considered in the design of Avionics products. Quantities for the Avionics market are typically very low when compared with commercial wireless applications, for example, the number of cell phones manufactured every single working day far exceeds the number of aircraft that are manufactured in the world in a year. Wireless systems for Avionics applications cover an extremely wide range in a number of dimensions, including frequency, system function, modulation type, bandwidth, and power. Owing to the number of systems aboard a typical aircraft, electromagnetic interference (EMI) and electromagnetic compatibility (EMC) between systems is a major concern, and EMI/EMC design and testing is a major factor in the flight certification testing of these systems.

13.1 Communications Systems, Voice and Data

VLF (very low frequency) is not used in civil aviation, but is used for low data rate transmission to and from strategic military platforms, owing to its ability to transmit reliably worldwide, including water penetration adequate to communicate with submerged submarines. These are typically very high power systems, with 100–200 kW of transmit power.

HF communications (high frequency), 2–30 MHz, have been used since the earliest days of aviation, and continues in use on modern aircraft, both civil and military. HF uses single sideband, suppressed carrier modulation, with a relatively narrow modulation bandwidth of about 2.5 kHz, typically several hundred-watt transmitters, and is capable of worldwide communications. Because propagation conditions vary with frequency, weather, ionosphere conditions, time of day, and sun spot activity, the whole HF-band will generally not be available for communications between any two points on earth at any given time, but some portion will be suitable. Establishing a usable HF link frequency between two stations is an essential part of the communications protocol in this band, and until recently, required a fair amount of time. Modern HF systems have automatic link establishment (ALE) software and hardware to remove that burden from the operator. Prior to the advent of communications satellites, HF was the only means of communicating with an aircraft for large parts of trans-oceanic flights.

VHF Communications (very high frequency) as used in the aviation world comprises two different frequency bands. The Military uses 30–88 MHz, with narrow band FM modulation, as an air to ground link for close air support, and for tactical communications on the ground. RF power for this usage ranges from 10 to 50 W. Civil and military aircraft both use the 116–136 MHz band with standard double sideband AM modulation for air traffic control (ATC) purposes, typically with 10–30 W of carrier power. While this band has been in use for decades, the technical requirements for radios in this band have not remained constant. The proliferation of potential interfering sites and competition for spectrum space has led to adapting ever narrower channel spacing and increased dynamic range requirements. An aircraft operating near an airport in an urban environment will frequently have to fly near a commercial FM station operating at several tens of kilowatts in the 88–108 MHz band, without compromising communications capability. Currently, VHF ATC radios operate with both 25 and 8.33 KHz channel spacing. In addition, this band is being used for addressing and reporting system (ACARS) and data link purposes, known as VDL (VHF Data Link) with differing modes of operation, depending on the required data and throughput required. The ACARS function, also known as VDL Mode A, uses an AM minimum shift keying (MSK) modulation at 2.4 KBPS. VDL Modes 2 and 3 are the highest data rate modes currently defined, achieving 31.5 kbps using a Differential 8 Phase Shift Keying (D8PSK) modulation scheme in a 25 KHz channel. These two modes differ in the networking scheme, with Mode 2 using a collision sense, multiple access (CSMA) network protocol, and Mode 3 using a time division multiple access (TDMA) network. Both VDL Modes 2 and 3 operate in an air to ground mode only. VDL Mode 4 utilizes a Gaussian frequency shift keying (GFSK) modulation at 19.5 KBPS in a self-organizing TDMA (STDMA) network.

UHF communications (ultra high frequency) as used in the aviation world actually bridge the VHF and UHF regions, operating from 225 to 400 MHz. Various radios in use here range in power output from 10 to 100 W of carrier power for FM, and 10–25 W of AM carrier power. This band is used for Military communications, and many waveforms and modulation formats are in use here, including AM, FM, and a variety of pulsed, frequency hopping, anti-jam waveforms using various protocols and encryption techniques.

SatCom, short for Satellite Communications, is used for aircraft data link purposes, as well as for communications by the crew and passengers. Passenger telephones use this service. Various satellites are used, including INMARSAT, and Aero-H and Aero-I, all operating at L-band.

Data Links are a subset of the overall communications structure, setup to transmit digital data to and from aircraft. They cover an extremely wide range of operating frequencies, data rates, and security features. Data Link applications may be a shared use of an existing radio, or use a specialized transceiver, dedicated to the application. Examples include the ARINC Communication ACARS which uses the existing VHF Comm radio on civil aircraft, and the Joint Tactical Information Distribution System (JTIDS), which is a secure, fast hopping L-band system that uses a dedicated transceiver.

Data link usage is an application area that has been dominated by military applications, but is now seeing an increasing number of commercial uses. Virtually all of the network management and spread spectrum methods now coming in to use for personal communications were originally pioneered by Military secure data link applications. Code division multiple access (CDMA), TDMA, bi-phase shift keying (BPSK), quadrature phase shift keying (QPSK), MSK, continuous phase shift keying (CPSM), Gaussian minimum shift keying (GMSK), various error detecting and correcting codes, along with interleaving and various forms of data redundancy have all found applications in Military data link applications. A great deal of effort has been devoted to classes of orthogonal code generation and secure encryption methods for use in these systems. It is common to find systems using a variety of techniques together to enhance their redundancy and anti-jam nature.

One system, for example, uses the following techniques to increase the anti-jam capabilities of the system. Messages are encoded with an error detecting and correcting code, to provide a recovery capability for bit errors. The message is then further encoded using a set of orthogonal CCSK code symbols representing a smaller number of binary bits, that is, 32 chips representing 5 binary bits, giving some level of CDMA spreading, which further increases the probability of successful decoding in the face of jamming. The result is then multiplied with a pseudo-random encryption code, further pseudo-randomizing the information in a CDMA manner. The resultant bit stream is then modulated on a carrier in a CPSM format, and transmitted symbol by symbol, with each symbol transmitted on a separate frequency in a pulsed, frequency hopped TDMA sequence determined by another encryption key. The message itself may also be parsed into short segments and interleaved with other message segments to time spread it still further, and repeated with different data interleaving during different time slots in the TDMA network arrangement to provide additional redundancy. Such systems have a high degree of complexity, and a large throughput overhead when operating in their most secure modes, but can deliver incredible robustness of message delivery in the face of jamming and/or spoofing attempts by an adversary. In addition, these systems are incredibly difficult to successfully intercept and decode, which is frequently just as important to the sender as the successful delivery of the message to its intended user.

13.2 Navigation and Identification Systems

Navigation and Identification functions comprise a large share of the Avionics aboard a typical aircraft. There are systems used for enroute navigation, including long range radio navigation (LORAN), automatic direction finder (ADF), global positioning system (GPS), global orbital navigation satellite system (GLONASS), distance measuring system (DME), and tactical air navigation (TACAN) system. There are also systems used for approach navigation, including marker beacon, VHF omni range (VOR), instrument landing system (ILS), microwave landing system (MLS), radar altimeter, ground collision avoidance system (GCAS), and GPS. Finally, there are systems used for identification of aircraft and hazards, including air traffic control radar beacon system (ATCRBS), Mode S, identification friend or foe (IFF), traffic collision avoidance system (TCAS), and Radar.

Commercial airborne radar systems are typically operated in either C-Band, around 5 GHz, or more commonly at X-band at 9.33 GHz. Their primary purpose is weather surveillance and hazard avoidance. Many use Doppler techniques with color-coded displays to quantify and clearly display the hazard level in a particular storm front, allowing the pilot to make better weather avoidance decisions. The operating principles of radar will not be discussed in this chapter.

Current trends in Avionics radio navigation systems mirror that of consumer systems in providing multiple capabilities in a single unit. They are referred to as multi mode receivers (MMR), and may combine the functions of VHF omnidirectional range (VOR), ILS, marker beacon, GPS, and MLS functions. They are installed as doubly or triply redundant systems, just as their predecessors are, but you are still only replicating one system, not several as before.

All instrument approach and landing aids assist the pilot in landing the aircraft safely in low visibility conditions. Low visibility conditions are broken down into three categories, with one having three subcategories. Each system is qualified to one of these categories. Each category defines the point in altitude and range at which a pilot must make a decision to land the plane, or exercise a missed approach, and try again or divert to an alternative airport. All ILS must have the highest order of decision integrity. Electronic failures are impossible to prevent entirely, but these systems are designed so that they cannot fail in a fashion that does not notify the pilot. Commercial aircraft installations also always include redundant system installations. Instrument approach and landing decision points break down as follows:

- Cat I—decision height of 200 feet (60 m) and a runway visual range of 1800 feet (550 m)
- Cat II—decision height of 100 feet (30 m) and a runway visual range of 1200 feet (350 m)
- Cat IIIa—decision height of 50 feet (15 m) and a runway visual range of 700 feet (200 m)
- Cat IIIb—decision height of 50 feet (15 m) and a runway visual range of 150 feet (50 m)
- Cat IIIc—decision height of 0 feet, (0 m) and a runway visual range of less than 150 feet (50 m)

Usually, Cat III operations are flown clear through to touchdown under autopilot control. It should be noted that, while Cat IIIc operations are well within the capabilities of current automatic flight control systems, relatively few airports have installed companion surface movement guidance systems that can accurately lead an aircraft to the terminal in near zero visibility, and consequently Cat IIIc landings are not allowed at these airports.

Many of the radio navigation systems rely on varying complexity schemes of spatial modulation for their function. In almost all cases, there are multiple navaids functional at a given airport ground station, including marker beacons, a DME or TACAN, a VOR, and an ILS and/or MLS system.

In a typical example, the VOR system operates at VHF (108–118 MHz) to provide an aircraft with a bearing to the ground station location in the following manner. The VOR transmits continuously at its assigned frequency, with two antennas (or the electronically combined equivalent), providing voice transmission and code identification to ensure that the aircraft is tracking the proper station. The identification is in Morse code, and is a 2 or 3 letter word repeated with a modulation tone of 1020 Hz. The transmitted signal from the VOR station contains a carrier that is amplitude modulated at a 30 Hz rate by a 30 Hz variable signal, and a 9960 Hz subcarrier signal that is frequency modulated between 10,440 and 9,480 Hz by a 30 Hz reference signal. The reference signal is radiated from a fixed omnidirectional antenna, and thus contains no time varying spatial modulation signal. The variable signal is radiated from a rotating (electrically or mechanically), semidirectional element driven at 1800 rpm or 30 Hz, producing a spatial AM modulation at 30 Hz. The phasing of the two signals is set to be in phase at magnetic North, 90° out of phase at magnetic East, 180° when South, and 270° when West of the VOR station. VOR receivers function by receiving both signals, comparing their phase, and displaying the bearing to the station to the pilot. In an area directly above a VOR station is an area that has no AM component. This cone-shaped area is referred to as the “cone of confusion,” and most VOR receivers contain delays to prevent “Invalid Data” flags from being presented during the brief fly over time.

The ILS contains three functions, localizer, glide slope, and marker beacon. Three receivers are used in this system: a VHF localizer, a UHF glide slope, and a VHF marker beacon. Their functions are described in the following paragraphs.

The marker beacon receiver functions to decode audio and provide signaling output to identify one of three marker beacons installed near the runway. The outer beacon is typically about 5 miles out, the middle approximately 3500 feet, and the inner just off the runway end. These beacons radiate a narrow beam width, 75 MHz signal in a vertical direction, and each has a different distinct modulation code so the receiver can identify which one it is flying over.

The localizer transmitter is located at the far end of the runway, and radiates two intersecting lobes in the 108–112 MHz frequency band along the axis of the runway, with the left lobe modulated at 90 Hz, and the right lobe modulated at 150 Hz. The installation is balanced so that the received signal will be equally modulated along the centerline of the runway. The total beam width is approximately 5° wide. The localizer receiver uses this modulation to determine the correct approach path in azimuth. This pattern also extends beyond the departure end of the runway, and this is called the back course. In the case of an aircraft flying an inbound back course, the sensed modulations will be reversed, and there are no marker beacons or glide slope signals available on the back course. In the event of a missed approach and go around, the side lobes of the localizer antenna pattern could present erroneous or confusing bearing data and validity flags, so most installations utilize a localizer go-around transmitter on each side of the runway

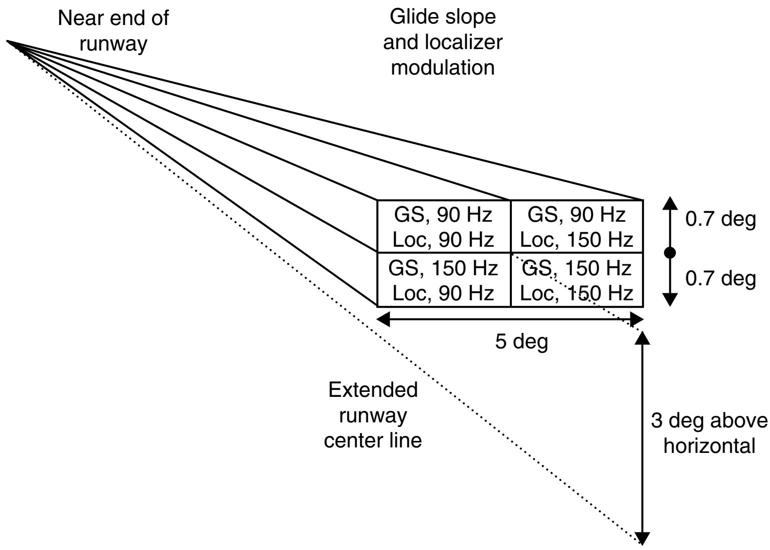


FIGURE 13.1 Glide slope and localizer modulation.

that radiates a signal directed outward from the side of the runway which is modulated with a go-around level modulation that identifies the signal and masks the sidelobe signals.

The glide slope transmitter is located close to the near end of the runway, and similarly radiates two intersecting lobes, one on top of the other, in the 329–335 MHz frequency band along the axis of, and at the angle of the desired approach glide slope, usually approximately 3°, with the upper lobe modulated at 90 Hz, and the lower lobe modulated at 150 Hz. The installation is balanced so that the received signal will be equally modulated along the centerline of the glide slope. The total beamwidth is approximately 1.4° wide. The localizer receiver uses this modulation to determine the correct glide slope path. The combination of localizer and glide slope modulations for the approach is illustrated in Figure 13.1. The back course, marker beacons, and go-around patterns are deleted for simplicity of visualization.

The addition of MLS technology to an MMR provides flight crews with an alternative to existing ILS capability wherever MLS is installed. MLS is currently installed, certified, and in use at several European airports, and more are planned in the near future. The military also uses MLS, because it can be made portable, and set up at an otherwise unimproved air strip in a matter of hours, not months, as is the case for a conventional VOR, ILS, and Localizer installation, and removed and reused at another time and place. MLS technology improves margins of safety around airports in highly developed areas, especially when weather conditions degrade and Category II or better capability is required to maintain airport capacity. MLS is used in conjunction with DME/P, which is the Distance Measuring Equipment, Precision to determine slant range measurements, while the MLS provides the angular data needed to determine present position relative to the runway. One of the major advantages that MLS has over the current VHF/UHF ILS system is the relative ease and cost of an installation. Because an ILS uses the ground in front of the glide slope antenna to form the beam, a large area must be level. Frequently, the cost of ground preparation for a VHF/UHF ILS site exceeds the equipment costs. In addition, a VHF/UHF ILS is sensitive to nearby reflecting surfaces, which can reflect and distort the glide slope beam. Consequently, aircraft and other vehicles must be held a long distance from the takeoff threshold to stay clear of the critical reflection area, and there are strict taxi restrictions imposed during periods of low visibility flight operations. MLS, operating at 5 GHz, does not suffer nearly as severely from these effects, and is much faster and easier to install and calibrate. Multipath distortions and errors at 5 GHz can be at high levels, and require much smaller reflective surfaces to create, but they also are varying over very small distances, allowing easier receiver designs to accommodate.

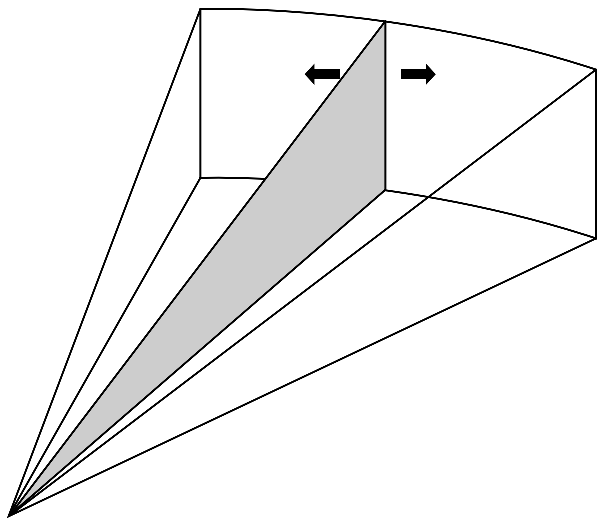


FIGURE 13.2 Microwave landing system azimuth scan.

MLS operates on 200 channels in a band from 5.03 to 5.09 GHz using a scanning beam in both elevation and azimuth to which the aircraft can use to determine its position relative to the desired glide slope. The MLS signal structure is a series of transmissions in sequence from the different antennas, and is very flexible, allowing Azimuth Elevation, and Back Azimuth to be transmitted in any order. The entire sequence takes approximately 75 ms to complete, and includes an azimuth scan, three elevation scans, and a back azimuth scan. Thus, the azimuth data are updated at $\approx$13.5 Hz, and elevation data at three times that, or $\approx$40 Hz. The MLS transmits a broad sector beam that contains a DPSK modulated preamble to provide the initial timing mark as well as information concerning the antenna's function and offset from the runway. MLS angle measurements are based on a time reference scanning beam (TRSB) system, which uses the timing between received pulses of a beam in both elevation and azimuth, that is scanned back and forth across the coverage area to extract angular data. In either case, the radiated beams are narrow, fan-shaped beams, scanned at 20,000° per second. The azimuth beam coverage is set at installation, and can cover up to $\pm 60°$. A typical installation is set to $\pm 40°$, while the minimum is $\pm 10°$. The azimuth coverage does not have to be symmetrical around the runway centerline. For example, it could be set to $+10°$, and $-40°$ to avoid mountains or some other obstruction. Figure 13.2 is a pictoral representation of the azimuth scanning beam format.

The elevation beam is set to scan from a minimum of 0.9° to a maximum of 15°. Typical fixed wing aircraft glide slopes on approach are around 3°. Figure 13.3 shows the scanning beam in elevation.

As stated above, the time difference between the received pulses on the "To" scan and the "Fro" scan is used to calculate the angle relative to the MLS transmitter and hence the airport centerline. Figure 13.4 shows the principle of TRSB angular measurement in azimuth, with elevation measured in the same manner.

Note that the scanned beam does not have to be modulated at all. The receiver uses amplitude information only to establish the timing between the pulses, and calculates the aircraft's angular position from the timing between received pulses. When this information is combined with slant range from the DME/P system, the aircraft's current position is fully established. Like the current ILS installations, there is a provision included to cover the back side of the runway for take offs, missed landings, and so on. Unlike the older VHF/UHF ILS system, which requires a straight in path down the glide slope, MLS can support arbitrarily curved approach paths. This is useful in maintaining separation distances in highly congested approach environments, and especially useful for the Military where mountains, structures, or enemy activity near the airfield might preclude a classic straight in glide path.

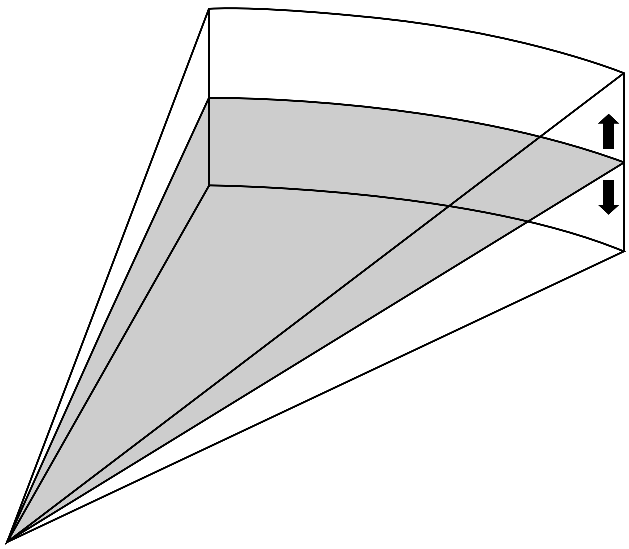


FIGURE 13.3 Microwave landing system elevations scan.

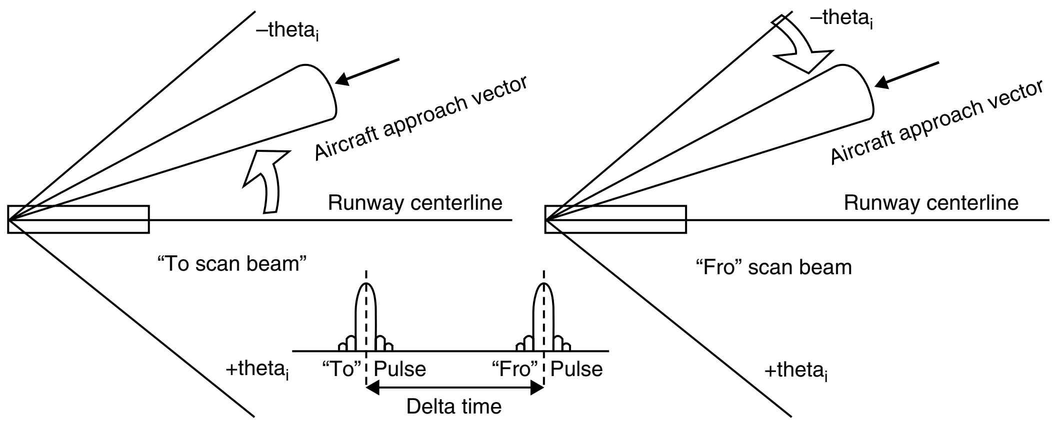


FIGURE 13.4 Microwave landing system angle calculation.

GPS is coming online as a viable approach navigation aid. The installation of wide area augmentation service (WAAS) and local area augmentation service (LAAS) has enhanced the accuracy of civil GPS to the point where it is now useful for approach navigation service. Both these services transmit corrections to GPS to allow significant improvements in position accuracy. These augmentations are also at higher levels, and not as easily jammed (deliberately or accidentally), as the very low-level GPS signal is. This and the advances in GPS receivers that track many satellites simultaneously, have allowed integrity monitoring to be built into the system that can meet the integrity requirements of approach navigation. In the United States, the FAA has selected GPS as the next generation ILS. There are currently MMR available that include GPS landing capability, but the airport infrastructure and subsequent certification to Cat I requirements is not expected before 2008 or 2009.

Another example of an Avionics System in widespread use that has significant RF content is the Traffic Alert and Collision Avoidance System, TCAS. TCAS functions in close harmony with the ATCRBS at 1030 and 1090 MHz, which has been in operation for several decades. The ATCRBS system comprises a system of ground-based interrogators and air borne responders, which serve to provide the ATC system with the necessary information to safely manage the air space.

There are several interrogation/reply modes incorporated into the ATCRABS/TCAS system. Mode A and Mode C interrogations use a combination of pulse position and width coding, and replies are in a pulse position/timing format. The Mode A response identifies the aircraft responding. Aircraft equipped with Mode C transponders interrogated by the ATC's Secondary Surveillance Radar system respond with a coded pulse train that includes altitude information, allowing automated altitude tracking by the ATC system. TCAS incorporates Mode C and Mode Select, or Mode S as it is more commonly known, which incorporates a data link capability, and interrogates these transponders in other aircraft. Mode S uses pulse position coding for interrogations and replies, and differential phase shift keying (DPSK) to send data blocks containing 56 or 112 data chips. All modes of transponder responses to an interrogation are designed to be tightly controlled in time so that slant range between the interrogator and responder are easily calculated. Altitude is directly reported in both Mode C and Mode S responses as well. Specific details of all of the allowable interrogation, reply, and data transmissions would take up far too much room to include here. As with most essential Avionics equipment, passenger aircraft will have two TCAS systems installed for redundancy.

TCAS provides for a surveillance radius of 30 miles or greater and displays detected aircraft in four categories to the pilot: Other, Proximate, TA, and RA. Each has its own unique coding on the display. Intruders classified as "Other" are outside of 6 nautical miles (nmi), or more than ± 1200 feet of vertical separation, and whose projected flight path does not take them closer than these thresholds. A "Proximate" aircraft is within 6 nmi, but whose flight path does not represent a potential conflict. TCAS provides the pilot with audible advisory messages for the remaining two categories; traffic advisories (TA) for "Proximate" aircraft when they close to within 20–48 s of closest point of approach (CPA) announced as "Traffic, Traffic," and resolution advisories (RA) for aircraft whose flight path is calculated to represent a threat of collision. RA are given 15–35 s prior to CPA, depending on altitude. RAs are limited to vertical advisories only; no turns are ever commanded. RAs may call for a climb, a descent, or to maintain or increase the rate of climb or descent. RAs are augmented by displaying a green arc on the vertical speed indicator, which corresponds to the range of safe vertical speeds.

TCAS antennas are multielement, multioutput port antennas, matched quite closely in phase and gain so that the receiver may determine the Angle of Arrival of a received signal. With altitude, slant range, and angle of arrival of a signal known, the present position, relative to own aircraft is computed. On the basis of changes between replies, the CPA is computed, and the appropriate classification of the threat is made. If both aircraft are equipped with Mode S capability, information is exchanged and RAs are coordinated to ensure that both aircraft do not make the same maneuver, requiring a second RA to resolve.

From an RF and Microwave point of view, there are several challenging aspects to a TCAS design.

1. The antenna requires four elements located in close proximity within a single enclosure that radiate orthogonal cardiod patterns matched in gain and phase quite closely, that is, 0.5 dB and 5°, typically. With the high amount of mutual coupling present, this presents a challenging design problem.
2. Because of the high traffic density that exists within the United States and Europe, the potential for multiple responses and interference between aircraft with the same relative angle to the interrogating aircraft but different ranges is significant. In order to prevent this, a "Whisper Shout" algorithm has been implemented for TCAS interrogations. This involves multiple interrogations, starting at low power, with each subsequent interrogation being one dB higher in power. This requires control of a high power (typically in excess of 1 kW) transmitter output over a 20 dB range, with better than 1 dB linearity. Attaining this linearity with high-speed switches and attenuators capable of handling this power level is not an easy design task. The Whisper Shout pulse sequence is illustrated in Figure 13.5.

Each interrogation sequence consists of four pulses, S1, P1, P3, and P4. Each pulse is 0.8 μS in width, with a leading edge spacing of 2.0 μS for pulses S1–P1 and P3–P4. There is a 21 μS spacing from P1 to P3 (compressed in the figure). Each interrogation sequence will be increased in power level in 1 dB steps over a 20 dB range. A transponder will only reply to pulse sequences that meet the timing requirements, and whose amplitude is such that S1 is below the receiver noise floor, and P1 is above a minimum trigger level.

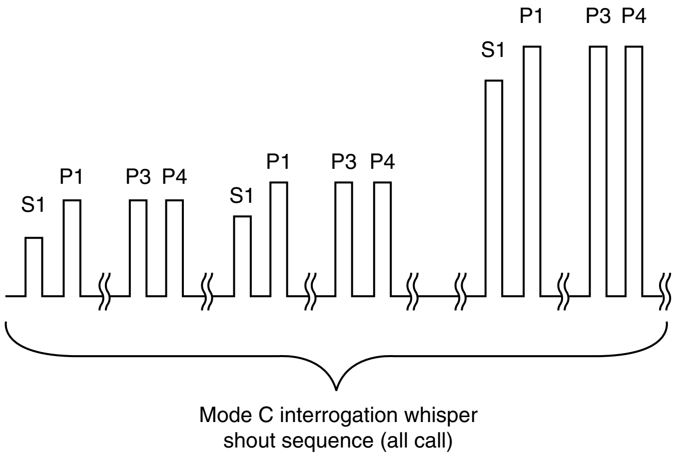


FIGURE 13.5 TCAS whisper shout pulse modulation.

This effectively limits the number of responses to the whisper shout sequence from any aircraft to about two, thus limiting the inter-aircraft interference.

13.3 Passenger Business and Entertainment Systems

Entertainment is the latest application of Microwave Technology to the aircraft industry. While the traditional definition of Avionics has referred to cockpit applications, in recent years there has been a move to receive Direct Broadcast TV services, and provide Internet Access services to passengers via individual displays at each seat. These applications require steerable, aircraft mounted antennas at X- or Ku-band capable of tracking the service provider satellites, low noise down-converters, some sort of closed-loop tracking system, typically utilizing the aircraft's inertial navigation system, and a receiver(s) with multiple tunable output channels, plus the requisite control at each seat to select the desired channel. Such a system is thus much more complex and expensive than a household Direct Broadcast System.

13.4 Military Systems

Military aircrafts utilize all of the avionics systems listed above (with the exception of entertainment), and depending on the mission of the individual airframe, may also contain systems to spoof, jam, or just listen to other users of any or all of those systems. The radar systems on military aircraft may be much more powerful and sophisticated than those required for civil use, and serve a variety of weapons or intelligence gathering functions in addition to the basic navigation/weather sensing functions of civil aircraft. In fact, virtually all of the functions used by civil aircraft may be exploited or denied for a military function, and systems exist on various platforms to accomplish just that.

There is a major initiative underway within the U.S. Department of Defense (DoD) to field a new generation of communications systems that are software defined radios, capable of operating with many different waveforms, depending on the software load (waveform) installed. The DoD has mandated that future radio systems in the 2–2000 MHz frequency range be designed to a software communications architecture (SCA). The DoD refers to this system as the joint tactical radio system (JTRS) and there are versions for ground vehicular, airborne, ship, handheld, manpack, transportable, and fixed ground currently under development. It is intended to integrate these communications systems into a Network Centric system, allowing much broader information sharing than is possible with current systems. Since these systems must function in the severe military environments, which include high dynamic range, high vibration, wide temperature ranges, and challenging jamming environments, design of RF hardware for these systems is highly challenging. Current developments are focused on a half dozen or so anti-jam and wide band networking waveforms, but future plans call for hosting of up to 30 different waveforms, including all of the legacy military usage, such as Single Sideband HF, VHF AM and FM, UHF AM and FM, several anti-jam waveforms with varying hopping and cryptographic requirements, data links, and some newly defined wide band networking waveforms. Required bandwidths range from 3 kHz to 30 MHz, depending on the frequency and waveform.

14 Continuous Wave Radar

Samuel O. Piper
James C. Wiltse
Georgia Tech Research Institute

Continuous wave (CW) radar employs a transmitter which is on all or most of the time. Unmodulated CW radar is very simple and is able to detect the Doppler-frequency shift in the return signal from a target, which has a component of motion toward or away from the transmitter. While such a radar cannot measure range, it is used widely in applications such as police radars, motion detectors, burglar alarms, proximity fuzes for projectiles or missiles, illuminators for semi-active missile guidance systems (such as the Hawk surface-to-air missile), and scatterometers (used to measure the scattering properties of targets or clutter such as terrain surfaces) (Komarov et al., 2003; Nathanson, 1991; Pace, 2003; Saunders, 1990; Ulaby and Elachi, 1990).

Modulated versions include frequency-modulated continuous wave (FMCW), interrupted frequency-modulated (IFMCW), and phase-modulated. Typical waveforms are indicated in Figure 14.1. Such systems are used in altimeters, Doppler navigators, proximity fuzes, over-the-horizon radar, and active seekers for terminal guidance of air-to-surface missiles. The term *continuous* is often used to indicate a relatively long waveform (as contrasted to pulse radar using short pulses) or a radar with a high duty cycle (for instance, 50% or greater, as contrasted with the typical duty cycle of $<10\%$ for the usual pulse radar). As an example of a long waveform, planetary radars may transmit for up to 10 h and are thus considered to be CW (Freiley et al., 1992). Another example is interrupted CW (or pulse-Doppler) radar, where the transmitter is pulsed at a high rate for 10–60% of the total time (Nathanson, 1991). All of these modulated CW radars are able to measure range.

The first portion of this chapter discusses concepts, principles of operation, and limitations. The latter portion describes various applications. In general, CW radars have several potential advantages over

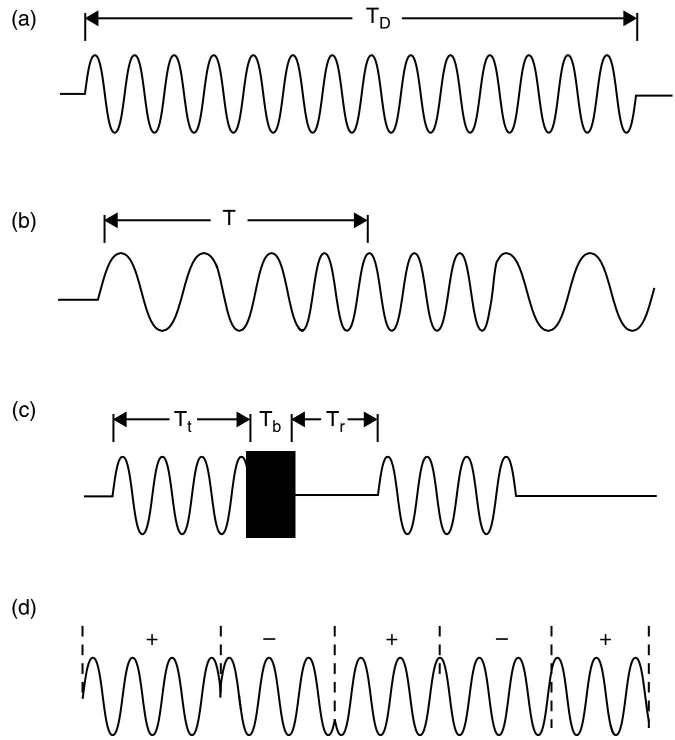


FIGURE 14.1 Waveforms for the general class of CW radar: (a) continuous sine wave CW, (b) frequency-modulated CW, (c) interrupted CW, and (d) binary phase-coded CW. (From F. E. Nathanson, *Radar Design Principles*, New York: McGraw-Hill, 1991, p. 450. With permission.)

pulse radars. Advantages include simplicity and the fact that the transmitter leakage is used as the local oscillator, transmitter spectral spread is minimal (not true for wide-deviation FMCW), and peak power is the same as (or only a little greater than) the average power. This latter situation means that the radar is less detectable by simple intercepting equipment.

The largest disadvantage for CW radars is the need to provide antenna isolation (reduce spillover) so that the transmitted signal does not interfere with the receiver. In pulse radar, the transmitter is off before the receiver is enabled (by means of a duplexer and/or receiver-protector switch). Isolation is frequently obtained in the CW case by employing two antennas, one for transmission and one for reception. When this is done, there is also a reduction of close-in clutter return from rain or terrain. A second disadvantage is the existence of noise sidebands on the transmitter signal, which reduce sensitivity because the Doppler frequencies are relatively close to the carrier. This is considered in more detail in the following sections.

14.1 CW Doppler Radar

If a sine wave signal were transmitted, the return from a moving target would be Doppler-shifted in frequency by an amount given by the following equation:

$$f_{\mathrm{d}} = \frac{2v_{\mathrm{r}}f_{\mathrm{T}}}{c} = \text{Doppler frequency} \tag{14.1}$$

where f_{T} is transmitted frequency; c is velocity of propagation, 3×10^8 m/s; and v_{r} is radial component of velocity between radar and target.

TABLE 14.1 Doppler Frequencies for Several Transmitted Frequencies and Various Relative Speeds (1 m/s = 2.237 mph)

Microwave Frequency—f_T (GHz)	Relative Speed			
	1 m/s (Hz)	300 m/s (kHz)	1 mph (Hz)	600 mph (kHz)
3	20	6	8.9	5.4
10	67	20	30	17.9
35	233	70	104	63
95	633	190	283	170

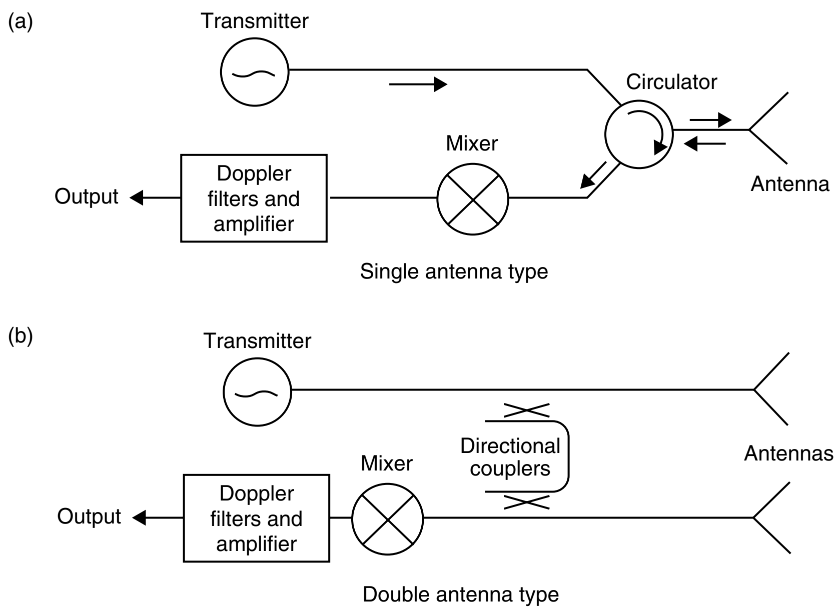


FIGURE 14.2 Block diagrams of CW-Doppler radar systems: (a) single antenna type and (b) double antenna type.

Using Equation 14.2 the Doppler frequencies have been calculated for several speeds and are given in Table 14.1.

As may be seen, the Doppler frequencies at 10 GHz (X-band) range from 30 Hz to about 18 kHz for a speed range between 1 and 600 mph. The spectral width of these Doppler frequencies will depend on target fluctuation and acceleration, antenna scanning effects, frequency variation in oscillators or components (e.g., due to microphonism from vibrations), but most significantly, by the spectrum of the transmitter, which inevitably will have noise sidebands that extend much higher than these Doppler frequencies, probably by orders of magnitude. At higher microwave frequencies the Doppler frequencies are also higher and more widely spread. In addition, the spectra of higher frequency transmitters are also wider, and, in fact, the transmitter noise-sideband problem is usually worse at higher frequencies, particularly at millimeter wavelengths (i.e., above 30 GHz). These characteristics may necessitate frequency stabilization or phase locking of transmitters to improve the spectra.

Simplified block diagrams for CW Doppler radars are shown in Figure 14.2. The transmitter is a single-frequency source and leakage (or coupling) of a small amount of transmitter power serves as a local oscillator signal in the mixer. This is called homodyning. The transmitted signal will produce a Doppler-shifted return from a moving target. In the case of scatterometer measurements, where, for example, terrain reflectivity is to be measured, the relative motion may be produced by moving the radar (perhaps on a vehicle) with respect to the stationary target (Wiltse et al., 1957). The return signal is collected by the antenna and then also fed to the mixer. After mixing with the transmitter leakage, a difference frequency will be produced, which is the Doppler shift. As indicated in Table 14.1, this difference is bound to range from low audio to over 100 kHz, depending on relative speeds and choice of microwave frequency. The Doppler amplifier and filters are chosen based on the information to be obtained, and this determines the amplifier bandwidth and gain, as well as the filter bandwidth and spacing. The transmitter leakage may include reflections from the antenna and/or nearby clutter in front of the antenna, as well as mutual coupling between antennas in the two-antenna case.

The detection range for such radar can be obtained from the following (Nathanson, 1991):

$$R^4 = \frac{P_T G_T L_T A_e L_R L_p L_a L_s \sigma_T}{(4\pi)^2 k T_s b(S/N)} \tag{14.2}$$

where

R is the detection range of the desired target;
P_T is the average power;
G_T is the transmit power gain of the antenna with respect to an omnidirectional radiator;
L_T is the losses between the transmitter output and free space including power dividers, waveguide or coax, radomes, and any other losses not included in A_e;
A_e is the effective aperture of the antenna, which is equal to the projected area in the direction of the target times the efficiency;
L_R is the receive antenna losses defined in a manner similar to the transmit losses;
L_p is the beam shape and scanning and pattern factor losses;
L_a is the two-way-pattern propagation losses of the medium, often expressed as $\exp(-2\alpha R)$, where α is the attenuation constant of the medium and the factor 2 is for a two-way path;
L_s is signal-processing losses that occur for virtually every waveform and implementation;
σ_T is the radar cross section of the object that is being detected; k is the Boltzmann's constant (1.38×10^{-23} W/HzK);
T_s is the system noise temperature; b is the Doppler filter or speedgate bandwidth;
S/N is signal-to-noise ratio;
S_{min} is the minimum detectable target-signal power that, with a given probability of success, the radar can be said to detect, acquire, or track in the presence of its own thermal noise or some external interference.

Since all these factors (including the target return itself) are generally noise-like, the criterion for a detection can be described only by some form of probability distribution with an associated probability of detection P_D and a probability that, in the absence of a target signal, one or more noise or interference samples will be mistaken for the target of interest.

While the Doppler filter should be a matched filter, it usually is wider because it must include the target spectral width. There is usually some compensation for the loss in detectability by the use of postdetection filtering or integration (Nathanson, 1991, p. 449).

The Doppler system discussed above has a maximum detection range based on signal strength and other factors, but it cannot measure range. The rate of change in signal strength as a function of range has sometimes been used in fuzes to estimate range closure and firing point, but this is a relative measure.

14.2 FMCW Radar

The most common technique for determining target range is the use of frequency modulation. Typical modulation waveforms include sinusoidal, linear sawtooth, or triangular, as illustrated in Figure 14.3. For a linear sawtooth, a frequency increasing with time may be transmitted. Upon being reflected from a stationary point target, the same linear frequency change is reflected back to the receiver, except it has a time delay, which is related to the range to the target. The time is $T = (2R)/c$ where R is the range. The

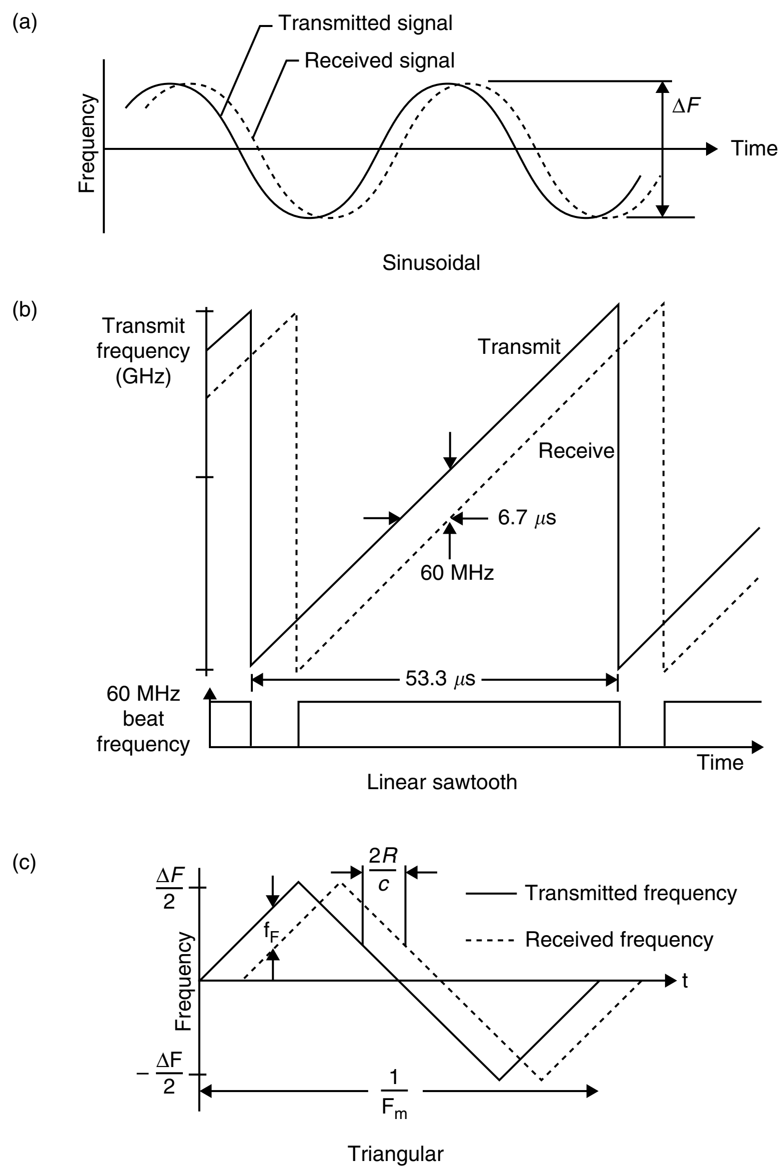


FIGURE 14.3 Frequency versus time waveforms for FMCW radar: (a) sinusoidal, (b) linear sawtooth, and (c) triangular modulations.

received signal is mixed with the transmit signal, and the difference or beat frequency (F_b) is obtained. (The sum frequency is much higher and is rejected by filtering.) For a stationary target, the beat frequency for a sawtooth waveform is given by

$$F_b = \frac{2R}{c} \cdot \Delta F \cdot F_m \tag{14.3}$$

and for a triangle wave, the beat frequency is

$$F_b = \frac{4R}{c} \cdot \Delta F \cdot F_m \tag{14.4}$$

where ΔF is the frequency deviation and F_m is the modulation rate.

The beat frequency is constant except near the turnaround region of the sawtooth, but, of course, it is different for targets at different ranges. (If it is desired to have a constant intermediate frequency for different ranges, which is a convenience in receiver design, then the modulation rate or the frequency deviation must be adjusted.) Multiple targets at a variety of ranges will produce multiple-frequency outputs from the mixer and are handled in the receiver by a bank of range-bin filters as may be provided by digital signal processing using fast Fourier transform techniques.

If the target is moving with a component of velocity toward (or away) from the radar, then there will be a Doppler frequency component added to (or subtracted from) the difference frequency (F_b), and the Doppler will be slightly higher at the upper end of the sweep range than at the lower end. This will introduce an uncertainty or ambiguity in the measurement of range, which may or may not be significant, depending on the parameters chosen and the application. For example, if the Doppler frequency is low (as in an altimeter) and/or the difference frequency is high, the error in range measurement may be tolerable. For the symmetrical triangular waveform, a Doppler less than F_b averages out, since it is higher on one-half of a cycle and lower on the other half. With a sawtooth modulation, only a decrease or increase is noted, since the frequencies produced in the transient during a rapid flyback are out of the receiver passband. Exact analyses of triangular, sawtooth, and dual triangular, dual sawtooth, and combinations of these with noise have been carried out by Tozzi (1972). Specific design parameters are given later in this chapter for an application utilizing sawtooth modulation in a missile terminal guidance seeker.

For the case of sinusoidal frequency modulation, the spectrum consists of a series of lines spaced away from the carrier by the modulation frequency or its harmonics. The amplitudes of the carrier and these sidebands are proportional to the values of the Bessel functions of the first kind ($J_n, n = 0, 1, 2, 3, \ldots$), whose argument is a function of the modulation frequency and range. By choosing a particular modulation frequency, the values of the Bessel functions and thus the characteristics of the spectral components can be influenced. For instance, the signal variation with range at selected ranges can be optimized, which is important in fuzes. A short-range dependence that produces a rapid increase in signal, greater than that corresponding to the normal range variation, is beneficial in producing well-defined firing signals. This can be accomplished by proper choice of modulation frequency and filtering to obtain the signal spectral components corresponding to the appropriate order of the Bessel function. In a similar fashion, spillover and/or reflections from close-in objects can be reduced by filtering to pass only certain harmonics of the modulation frequency (F_m). Receiving only frequencies near $3F_m$ results in considerable spillover rejection but at a penalty of 4–10 dB in signal-to-noise ratio (Nathanson, 1991).

For the sinusoidal modulation case, Doppler frequency contributions complicate the analysis considerably. For details of this analysis, the reader is referred to Nathanson (1991) or Saunders (1990).

14.3 Interrupted Frequency-Modulated CW

To improve isolation during reception, the interrupted frequency-modulated CW (IFMCW) format involves preventing transmission for a portion of the time during the frequency change. Thus, there are frequency gaps, or interruptions, as illustrated in Figure 14.4. This shows a case where the transmit time equals the round-trip propagation time, followed by an equal time for reception. This duty factor of 0.5 for the waveform reduces the average transmitted power by 3 dB relative to using an uninterrupted transmitter. However, the improvement in the isolation should reduce the system noise by more than 3 dB, thus improving the signal-to-noise ratio (Piper, 1987). For operation at short range, Piper (1987) points out that a high-speed switch is required. He also points out that the ratio of frequency deviation to beat frequency should be an even integer and that the minimum ratio is typically 6, which produces an out-of-band loss of 0.8 dB (see also Brooker, 2005).

IFMCW may be compared with pulse compression radar if both use a wide bandwidth. Pulse compression employs a "long" pulse (i.e., relatively long for a pulse radar) with a large frequency deviation or "chirp." A long pulse is often used when a transmitter is peak-power limited, because the longer pulse produces more energy and gives more range to targets. The frequency deviation is controlled in a

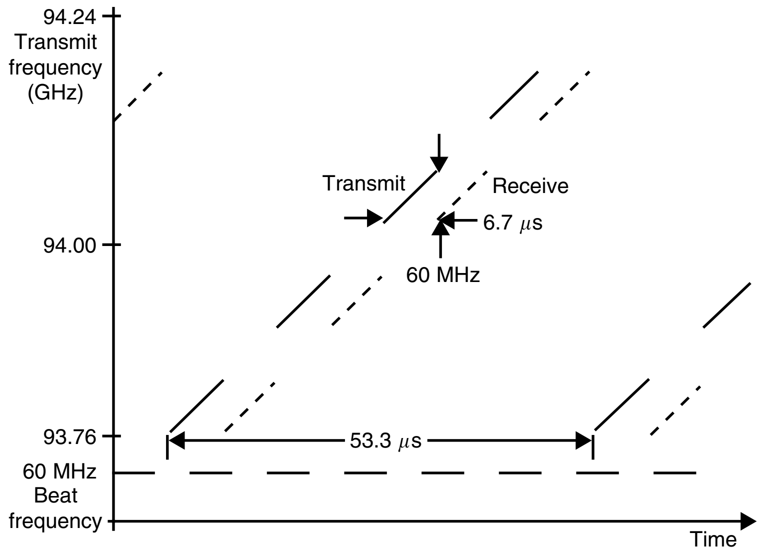


FIGURE 14.4 Interrupted FMCW waveform. (From S. O. Piper, MMW seekers, in *Principles and Applications of Millimeter Wave Radar*, N. C. Currie and C. E. Brown (Eds.), Norwood, MA: Artech House, 1987, p. 683, chap. 14. With permission.)

predetermined way (frequently a linear sweep) so that a matched filter can be used in the receiver. The large time-bandwidth product permits the received pulse to be compressed in time to a short pulse in order to make an accurate range measurement. A linear-sawtooth IFMCW having similar pulse length, frequency deviation, and pulse repetition rate would thus appear similar, although arrived at from different points of view.

14.4 Applications

Space does not permit giving a full description of the many applications mentioned at the beginning of this chapter, but several will be discussed.

14.4.1 Radar Proximity Fuzes

Projectiles or missiles designed to be aimed at ships or surface land targets often need a height-of-burst (HOB) sensor (or target detection device) to fire or fuze the warhead at a height of a few meters. There are two primary generic methods of sensing or measuring height to generate the warhead fire signal. The most obvious, and potentially the most accurate, is to measure target round-trip propagation delay employing conventional radar ranging techniques. The second method employs a simple CW Doppler radar or variation thereof, with loop gain calibrated in a manner that permits sensing the desired burst height by measurement of target return signal amplitude and/or rate of change. Often the mission requirements do not justify the complexity and cost of overcoming the short range eclipsing challenges associated with the radar ranging approach. Viable candidates are thus narrowed down to variations on the CW Doppler fuze.

In its simplest form, the CW Doppler fuze consists of a fractional watt RF oscillator, homodyne detector, Doppler amplifier, Doppler envelope detector, and threshold circuit. When the Doppler envelope amplitude derived from the returned signal reaches the preset threshold, a fire signal is generated. The height at which the fire signal occurs depends on the radar loop gain, threshold level, and target reflectivity. Fuze gain is designed to produce the desired height of burst under nominal trajectory angle and target reflectivity conditions, which may have large fluctuations due to glint effects. Deviations from the desired height due to antenna gain variations with angle, target reflectivity, and fuze gain tolerances are accepted. A loop gain change of 6 dB (2 to 1 in voltage), whether due to a change in target reflection coefficient, antenna gain, or whatever, will result in a 2 to 1 HOB change.

HOB sensitivity to loop gain factors can be reduced by utilizing the slope of the increasing return signal, or so-called rate-of-rise. Deriving HOB solely from the rate-of-rise has the disadvantage of rendering the fuze sensitive to fluctuating signal levels, such as might result from a scintillating target. The use of logarithmic amplifiers decreases the HOB sensitivity to the reflectivity range. An early (excessively high) fire signal can occur if the slope of the signal fluctuations equals the rate-of-rise threshold of the fuze. In practice, a compromise is generally made in which Doppler envelope amplitude and rate-of-rise contribute in some proportion of HOB.

Another method sometimes employed to reduce HOB sensitivity to fuze loop gain factors and angle of fall is the use of FM sinusoidal modulation of suitable deviation to produce a range correlation function comprising the zero order of a Bessel function of the first kind. The subject of sinusoidal modulation is quite complex, but has been treated in detail by Saunders (1990, pp. 14.22–14.46 and 14.41). The most important aspects of fuze design have to do with practical problems such as low cost, small size, ability to stand very high-g accelerations, long life in storage, and countermeasures susceptibility.

14.4.2 Police Radars

Down-the-road police radars, which are of the CW Doppler type, operate at 10.525 GHz (X-Band), 24.150 GHz (K-Band), or in the 33.4–36.0 GHz (Ka-band) range, frequencies approved in the United States by the Federal Communications Commission. Antenna half-power beamwidths are typically in the 0.21–0.31 radian (12°–18°) range. The sensitivity is usually good enough to provide a range exceeding 800 m. Target size has a dynamic range of 30 dB (from smallest cars or motorcycles to large trucks). This means that a large target can be seen well outside the antenna 3-dB point at a range exceeding the range of a smaller target near the center of the beam. Thus, there can be uncertainty about which vehicle is the target. Fisher (1992) has given a discussion of a number of the limitations of these systems, but in spite of these factors, probably tens of thousands have been built.

The transmitter is typically a Gunn oscillator in the 30–100 mW power range, and antenna gain is usually around 20–24 dBi, employing circular polarization. The designs typically have three amplifier gains for detection of short, medium, or maximum range targets, plus a squelch circuit so that sudden spurious signal will not be counted. Provision is made for calibration to assure the accuracy of the readings. Speed resolution is about 1 mph. The moving police radar system uses stationary (ground) clutter to derive patrol car speed. Then closing speed, minus patrol car speed, yields target speed.

The limitations mentioned about deciding which vehicle is the correct target have led to the development of laser police radars, which utilize much narrower beamwidth, making target identification much more accurate. Of course, the use of microwave and laser radars has spawned the development of automotive radar detectors, which are also in wide use.

14.4.3 Altimeters

A very detailed discussion of FMCW altimeters has been given by Saunders (1990, pp. 14.34–14.36), in which he has described commercial products built by Bendix and Collins. The parameters will be summarized in the following text and if more information is needed, the reader may want to turn to other references (Bendix Corp., 1982; Maoz et al., 1991; Saunders, 1990; Stratahos, 2000). In his material, Saunders gives a general overview of altimeters, all of which use wide-deviation FM at a low modulation frequency. He discusses the limitations on narrowing the antenna pattern, which must be wide enough to accommodate attitude changes of the aircraft. Triangular modulation is used, since for this waveform, the Doppler averages out, and dual antennas are employed. There may be a step error or quantization in height (which could be a problem at low altitudes), due to the limitation of counting zero crossings. A difference of one zero crossing (i.e., 0.5 Hz) amounts to ±0.75 m for a frequency deviation of 100 MHz. At 100 MHz,

TABLE 14.2 Parameters for Two Commercial Altimeters

Manufacturer and Model	Modulation Frequency	Frequency Deviation	Prime Power	Weight (pounds)	Radiated Power
Bendix ALA-52A	150 Hz	130 MHz	30 W	11[a]	
Collins ALT-55	100 kHz	100 MHz		8	350 mW

[a] Not including antenna and indicator.

half-wavelength, the distance between zero crossings, is 1.5 m, which corresponds to an error of ±0.75 m. Irregularities are not often seen, however, since meter response is slow. In addition, if terrain is rough, there will be actual physical altitude fluctuations. Table 14.2 shows some of the altimeters' parameters.

These altimeters are not acceptable for military aircraft, because their relatively wide-open front ends make them potentially vulnerable to electronic attack or electronic countermeasures (ECM). A French design has some advantages in this respect by using a variable frequency deviation, a difference frequency that is essentially constant with altitude, and a narrowband front-end amplifier (Saunders, 1990).

14.4.4 Doppler Navigators

Radar Doppler navigators employ multiple downward-oriented antenna beams and they estimate relative velocity via the Doppler frequency shift in each of the antenna beams. The velocity component is mathematically integrated to estimate position for dead reckoning navigation. These systems are mainly sinusoidally modulated FMCW radars employing four separate downward looking beams aimed at about 15° off the vertical. Because commercial airlines have shifted to nonradar forms of navigation, these units are designed principally for helicopters. Saunders (1990) cites a particular example of a commercial unit operating at 13.3 GHz, employing a Gunn oscillator as the transmitter, with an output power of 50 mW, and utilizing a 30-kHz modulation frequency. A single microstrip antenna is used. It is a low-altitude equipment (below 15,000 ft) and the unit weighs less than 12 pounds. A second unit cited has an output power of 300 mW, dual antennas, dual modulation frequencies, and an altitude capability of 40,000 ft. Helicopter Doppler navigators include the AN/ASN-128 and the AN/ASN-157 (Buell, 1980, 1993).

14.4.5 Phase-Modulated CW Radar

An example of phase-modulated CW radar is a developmental obstacle avoidance radar described by Honeywell. This 35-GHz radar is derived from a 4.3 GHz high-duty cycle biphase modulated CW covert radar altimeter waveform for helicopter low-level collision avoidance (Proctor, 1997). This obstacle avoidance radar system, developed by NASA Ames and Honeywell, has a 20° vertical by 50° horizontal field of view. Figure 14.5 shows the radar on a helicopter.

This radar successfully detected metal high-tension towers and cables, water towers and small radio towers, wood telephone poles, vehicles, boats, fences, leafless trees, and ground return at less than 4° grazing angle (Raymer and Weingartner, 1994).

The radar requirements include 914.4 m (3000 ft) maximum range in 16 mm/h rainfall rate and for −20 dB surface reflectivity for 5° grazing angle. The maximum velocity is 150 knots and the false alarm rate is 5%. Low cost and low probability of intercept (LPI) and detection were also required (Becker and Almsted, 1995). Table 14.3 lists the radar parameters.

As described in Table 14.4, the 32 ns biphase code chip corresponds to 4.8 m (16 ft) range resolution. This will require a 31.3 MHz phase modulator and an analog-to-digital converter (ADC) sample rate of at least 62.5 MHz along with 62.5 MHz memory. A code sequence of at least 190 chips is required for 0.9 km maximum range. For a maximal length code of 190, the range sidelobe levels will be approximately 23 dB. The radar receiver must perform a 190-element correlation. For a brute force correlator, this will require 190 multiplies every 16 ns or 11.9×10^9 multiplies per second.

FIGURE 14.5 Honeywell 35-GHz biphase modulated CW obstacle avoidance radar [http://ccf.arc.nasa.gov/dx/basket/pix/RASCAL.jpg] (Photo—Dominic Hart).

TABLE 14.3 Honeywell Phase-Modulated CW Radar Parameters

Parameter	Value
RF center frequency	35 GHz
Transmit power	35 mW
Receiver noise figure	6 dB
Antenna gain	34 dBi
Beamwidth	3°
Antenna sidelobes	25 dB
Azimuth field of view	±45°
Elevation field of view	±10°
Bi-Phase code chip	32 ns or 4.8 m (16 ft)
Codes transmitted	1, 5, 7, 11, and 13 bit
Receiver bandwidth	25 kHz

TABLE 14.4 Honeywell Phase-Modulated CW Radar Waveform

Biphase Modulation Parameter	Performance	Requirements
32 ns Chip length (1,120 35 GHz cycles per chip)	4.8 m (16 ft) range resolution	31.3 MHz bandwidth phase modulator 62.5 MHz ADC sample rate: 8- to 10-bit (48–60 dB) dynamic range (ADC technology limitation) 62.5 MHz memory
190 Code length (190 code length $\times$32 ns chip length $\sim$6 μs code repetition interval)	0.9 km (3 kft) Unambiguous range	190 Element correlator 190 Multiplies every 16 ns 11.9×10^9 multiplies per second

TABLE 14.5 Phase-Modulated CW Challenges

To Improve Performance	Benefit	Requirements
Decrease chip length (analogous to compressed pulse length of pulse radar)	• Finer range resolution	• Wider bandwidth phase modulator • Faster ADC sample rate—reduces dynamic range (ADC technology limitation) • Faster memory
Increase code length Code length × chip length (analogous to pulse repetition interval of pulse radar)	• Longer unambiguous range • Lower range sidelobe levels	• Longer correlator • More operations

TABLE 14.6 PILOT Mk3 FMCW Radar Parameter Summary

	Pilot
RF center frequency	9.3 GHz, X(I)-band
Frequency agility bandwidth	400 MHz
Features	Reflected power canceller
Output power	1, 10, 100, and 1000 mW
Receiver noise figure	5 dB
Instrumented range	4.4, 11.1, 44.4 km (2.4, 6, 24 nautical miles)
Range cell size	2.4, 6, 24 m

This example illustrates the challenges for fine-resolution phase-modulated CW radar. As described in Table 14.5, in order to achieve finer range resolution, shorter chip length is required, which will require a corresponding increase in the phase modulator bandwidth and the ADC sample rate along with the memory. Longer unambiguous range or lower range sidelobe levels will require longer code sequences requiring a longer correlator with more operations.

14.4.6 PILOT FMCW Radars

The PILOT radar from Saab Bofors Dynamics was first developed in the 1980s. Shown in Figure 14.6, it is a navigation and detection radar for use on ships and submarines that employs the FMCW waveform to achieve low probability of intercept (LPI) performance. Table 14.6 lists the parameters of the PILOT radar. The PILOT Mk3 version includes frequency agility over the 400-MHz bandwidth from 9.1 to 9.5 GHz to enhance its LPI performance and a Reflected Power Canceller to permit the system to operate with standard single navigation radar antennas. The PILOT transmit power of 1 W gives it the same average power and detection range of pulsed navigation radars, but it enjoys an advantage relative to simple intercept receivers and antiradiation missile receivers that are not matched to the FMCW waveform. In addition, the operator can reduce the transmit power in 10 dB increments down to 1 mW for operation at shorter ranges and against larger targets (Saab, 2005) (Figure 14.6).

Lin et al. give analytical expressions and experimental results for canceling front-end reflections in FMCW radar. They determined that for a 2-GHz total frequency deviation and 10-ms modulation period, with 200-MHz/ms sweeping rate, and a 1-ns RF delay mismatch, in order to achieve 35-dB cancellation, which requires phase error less than 1°, the cancellation loop response time should be less than 13.8 ms, which corresponds to 73.5-kHz bandwidth. This requires a real-time cancellation control loop (Lin and Wang, 2004; Lin et al., 2004).

FIGURE 14.6 PILOT FMCW radar transceiver, signal processor, and remote control panel on left and transceiver interior on right (Saab, 2005).

14.4.7 Frequency Shift Keying CW Radar

CW radar can also use frequency modulation that steps back and forth between two frequencies. Eaton VORAD radars use this frequency shift keying (FSK) CW waveform. One version operated at 24.725-GHz RF center frequency dwelling approximately 5 ms at each of two frequencies separated by 500-kHz, which yields 300-m maximum unambiguous range. The radar had 0.5-mW transmit power, and 4.0° azimuth and 5.5° elevation beamwidths (Will, 1995).

The Eaton VORAD model EVT-300 Collision Warning System operates at 24.725 GHz with 3 mW (typical) transmitted RF power. The operating range is from 0.9 m to 107 m (3–350 ft) with 0.5–120 mph host vehicle speeds and 0.25–100 mph vehicle closing rates. Driver warnings are issued via beeper and LEDs. The "SmartCruise" add-on offers the adaptive cruise control (ACC) function (Eaton, 2005).

14.4.8 Millimeter-Wave Seeker for Terminal Guidance Missile

Terminal guidance for short-range (less than 2 km) air-to-surface missiles has seen extensive development in the last 2 decades. Targets such as tanks are frequently immersed in a clutter background, which may give a radar return that is comparable to that of the target. To reduce the clutter return in the antenna footprint, the antenna beamwidth is reduced by going to millimeter wavelengths. For a variety of reasons, the choice is usually a frequency near 35 or 90 GHz. Antenna beamwidth is inversely proportional to frequency; so in order to get a reduced beamwidth, we would normally choose 90 GHz; however, more deleterious effects at 90 GHz due to atmospheric absorption and scattering can modify that choice. In spite of small beamwidths, the clutter is a significant problem, and in most cases, signal-to-clutter is a more limiting condition than signal-to-noise in determining range performance. Piper (1987) analyzed the situation for 35- and 90-GHz pulse radar seekers and compared them with a 90-GHz FMCW seeker. His FMCW results will be summarized in the following text.

In his approach to the problem, Piper gives a summary of the advantages and disadvantages of a pulse system compared to the FMCW approach. One difficulty for the FMCW can be emphasized here. That is the need for a highly linear sweep, and, because of the desire for the wide bandwidth, this requirement is accentuated. Direct Digital Chirp Synthesizers (DDCS) offer high-fidelity linear FM sweeps over wide bandwidths. The STEL-2375B DDCS and the STEL-9949 DDCS from ITT Microwave offer 400-MHz bandwidth with 0.23-Hz frequency steps (ITT Microwave, 2004). Analog Devices offers the model AD9858 DDS featuring a 10-bit DAC operating up to 1 GSPS (Giga samples per second). The AD9858 is capable of generating a frequency-agile analog output sine wave of up to 400+ MHz (Analog Devices, 2006). The wide bandwidth is desired in order to average the clutter return and to smooth the glint effects. In particular, glint occurs from a complex target because of the vector addition of coherent signals scattered back to the receiver from various reflecting surfaces. At some angles, the vectors may add in phase (constructively) and at others they may cancel, and the effect is specifically dependent on wavelength. For a narrowband system, glint may provide a very large signal change over a small variation of angle, but, of course, at another wavelength it would be different. Thus, very wide bandwidth is desirable from this smoothing point of view, and typical numbers used in millimeter-wave radars are in the 450–650-MHz range. Piper chose 480 MHz.

Another trade-off involves the choice of FM waveform. Here the use of a triangular waveform is undesirable because the Doppler frequency averages out and Doppler compensation is then required. Thus, the sawtooth version is chosen, but because of the large frequency deviation desired, the difficulty of linearizing the frequency sweep is made greater. In fact, many components must be extremely wideband, and this generally increases cost and may adversely affect performance. On the other hand, the difference frequency (F_b) and/or the intermediate frequency (F_{IF}) will be higher and thus further from the carrier, so the phase noise will be lower. After discussing the other trade-offs, Piper chose 60 MHz for the beat frequency. A modern FMCW system is likely to use a lower beat frequency to reduce the ADC sample rate requirement for digital signal processing.

With a linear FMCW waveform, the inverse of the frequency deviation provides the theoretical time resolution, which is 2.1 ns for 480 MHz (or range resolution of 0.3 m). For an RF sweep linearity of 300 kHz, the range resolution is actually 5 m at the 1000-m nominal search range. (The system has a mechanically scanned antenna.) An average transmitting power of 25 mW was chosen, which was equal to the average power of the 5-W peak IMPATT assumed for the pulse system. The antenna diameter was 15 cm. For a target radar cross section of $20\,\mathrm{m}^2$ and assumed weather conditions, the signal-to-clutter and signal-to-noise ratios were calculated and plotted for ranges up to 2 km and for clear weather or 4-mm-per-hour rainfall. The results show that for 1-km range, the target-to-clutter ratios are higher for the FMCW case than the pulse system in clear weather or in rain, and target-to-clutter is the determining factor.

14.4.9 Automotive CW Radar

CW waveforms and particularly the FMCW waveform are attractive for automotive applications. The United States, Europe, and Japan have designated the 76–77 GHz band for automotive sensors. Daimler Chrysler introduced MMW radar for adaptive cruise control (ACC) in 1998 and other manufacturers including Nissan, Jaguar, BMW, Honda, and Toyota have followed with increasingly capable systems. Delphi Corporation offers the Forewarn® Smart Cruise Control with Headway Alert and Stop-and-Go. This ACC uses a mechanically scanning, 76-GHz FMCW, long-range radar sensor to detect objects in the vehicle's path up to 152 m ahead. Using the vehicle's braking and throttle systems, Smart Cruise Control automatically manages vehicle speed to maintain a time gap (following distance) set by the driver (Delphi Corporation, 2006). The European RadarNet project (http://www.radarnet.org/start.htm) is developing a low-cost radar network for automotive applications such as collision avoidance, collision warning, stop and go functionality, airbag precrash warning, parking aid, and so on. The radar network will consist of four short-range sensors that are synchronized at RF and one enhanced long-range sensor. The network uses a FMCW radar waveform (Chabert, 2004).

Quinstar Technology, Inc., describes a 76.5 GHz FMCW prototype radar for automobile collision warning. Output power options were from 10 mW to greater than 100 mW and sweep linearity was less than 0.1% over 400-MHz frequency deviation (Quinstar Technology, Inc., 2006). Fujitsu-Ten Company describes their single-chip MMIC 76-GHz FMCW automobile radar along with radars from Delphi, Bosch, Honda elesys, Denso, and Hitachi (Yamano, 2004). Table 14.7 summarizes these CW automotive radars. Itoh and Honjo (2005) reviewed recent advances in 77-GHz MMIC module design techniques for automotive radar applications requiring low cost, mass production, high-yield manufacturing, and

TABLE 14.7 Comparison of Automotive Radars Using the CW Waveform

Manufacturer	Delphi	Honda elesys	Denso	Hitachi	Fujitsu Ten
External dimensions (mm)	137 × 67 × 100	123 × 98 × 79	77 × 107 × 53	80 × 108 × 64	89 × 107 × 86
Modulation method	FMCW	FMCW	FMCW	Two-frequency CW	FMCW
Detection range	Approx. 1–150 m	4–100 m or greater	Approx. 2–150 m	Approx. 1–150 m	4–120 m or greater
Horizontal detection angle	Approx. ±5°	±8°	±10°	±8°	±8°
Angle detection method	Mechanical scan	Beam conversion	Phased array	Monopulse	Mechanical scan
EHF device	GUNN	MMIC	MMIC	MMIC	MMIC

testing. Another indicator of the maturation of this radar application is Anritsu's offering the model ME7220A Radar Test System (RTS) for automotive radars in the 76 to 77-GHz band with up to 300-MHz instantaneous bandwidth (Anritsu Company, 2006).

14.5 Summary Comments

From this brief review, it is clear that there are many uses for CW radars and various types (such as fuzes) have been produced in large quantities. Because of their relative simplicity, today there are continuing trends toward the use of digital processing and integrated circuits. In fact, this is exemplified in articles describing FMCW radars built on single microwave integrated circuit chips (Chang et al., 1995; Haydl et al, 1999; Maoz et al., 1991; Menzel, 1999).

Defining Terms

Doppler-frequency shift: The observed frequency change between the transmitted and received signal produced by motion along a line between the transmitter/receiver and the target. The frequency increases if the two are closing and decreases if they are receding.
Missile terminal guidance seeker: Located in the nose of a missile, a small radar with short-range capability that scans the area ahead of the missile and guides it during the terminal phase toward a target such as a tank.
Pulse Doppler: A coherent radar, usually having high pulse repetition rate and duty cycle and capable of measuring the Doppler frequency from a moving target. Has good clutter suppression and thus can see a moving target in spite of background reflections.

References

1. Analog Devices, *AD9858 Data Sheet*, 2006 http://www.analog.com/UploadedFiles/Data_Sheets/37263106237625AD9858_a.pdf
2. Anritsu Company, *ME 7220A Radar Test System (RTS) Brochure*, 2006, http://www.eu.anritsu.com/files/11410-00267.pdf
3. R. C. Becker and L. D. Almsted, Flight test evaluation of a 35 GHz forward looking altimeter for terrain avoidance, *IEEE AES Systems Magazine*, Vol. 10, 1995, pp. 19–22.
4. Bendix Corporation, *Service Manual for ALA-52A Altimeter*, Design Summary for the ALA-52A, Bendix Corporation, Ft. Lauderdale, Fla., May 1982.
5. G. M. Brooker, D. Birch, and J. Solms, W-Band airborne interrupted frequency modulated CW imaging radar, *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 41, No. 3, 2005, pp. 955–972.
6. H. Buell, Div. Kearfott, Co. Singer, and N. J. Wayne, Doppler radar systems for helicopters, Navigation, *Journal of the Institute of Navigation*, Vol. 27, No. 2, Summer 1980, pp. 124–131.
7. H. Buell and D. Doremus, GEC-Marconi Electron. Syst. Corp., Wayne, NJ, The AN/ASN-157; a single LRU Doppler Navigation System for helicopters, *Proceedings of the 12th AIAA/IEEE Digital Avionics Systems Conference* (*DASC*), 25–28 October 1993, pp. 260–266.
8. M. Chabert, C. Mailhes, N. Thomas, A. Ducasse, P. Michel, and F. Castani, On the use of high resolution spectral analysis methods in radar automotive, *Proceedings of the 1st International Workshop on Intelligent Transportation* (*WIT 2004*), 23–24 March 2004, Hamburg, Germany.
9. K. W. Chang, H. Wang, G. Shreve, J. G. Harrison, M. Core, A. Paxton, M. Yu, C. H. Chen, and G. S. Dow, Forward-looking automotive radar using a W-band single-chip transceiver, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 43, 1995, pp. 1659–1668.
10. Collins (Rockwell International), *ALT-55 Radio Altimeter System: Instruction Book*, 1984, Cedar Rapids, IA.
11. Delphi Corporation, *Delphi Forewarn®Smart Cruise Control with Headway Alert & Stop-and-Go*, 2005, http://www.delphi.com/pdf/ppd/safesec/for_stopgo.pdf

12. Eaton, 2005, http://www.truck.eaton.com/na/spec_products/product_features_specs/vorad/service.htm
13. P. D. Fisher, Improving on police radar, *IEEE Spectrum*, Vol. 29, 1992, pp. 38–43.
14. A. J. Freiley, B. L. Conroy, D. J. Hoppe, and A. M. Bhanji, Design concepts of a 1-MW CW X-band transmit/receive system for planetary radar, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 40, June 1992, pp. 1047–1055.
15. W. H. Haydl et al., Single-chip coplanar 94 GHz FMCW radar sensors, *IEEE Microwave and guided Wave Letters*, Vol. 9, February 1999, pp. 73–75.
16. Y. Itoh and K. Honjo, 77-GHz MMIC module design techniques for automotive radar applications, special section on recent technologies on devices, circuits, and systems for millimeter-wave ITS applications—review paper, *EICE Transactions on Electronics*, Vol. E88-C, No. 10, 2005, pp. 1939–1946, http://www.ietele.oxfordjournals.org/cgi/content/refs/E88-C/10/1939.
17. ITT Microwave, http://www.ittmicrowave.com/ddsModules.asp?subsection=0, 2004 and http://www.ittmicrowave.com/ddsModules.asp?subsection=1, 6 June 2004.
18. I. V. Komarov, S. M. Smolskiy, and D. K. Barton (translator), *Fundamentals of Short-Range FM Radar*, Norwood, MA: Artech House, September 2003.
19. K. Lin, R. H. Messerian, and Y. Wang, A digital leakage cancellation scheme for monostatic FMCW radar, *IEEE MTT-S International Microwave Symposium Digest* (IEEE Cat. No. 04CH37535), Vol. 2, 6–11 June 2004, pp. 747–750.
20. K. Lin and Y. E. Wang, Real-time DSP for reflected power cancellation in FMCW radars, *Proceedings of the Vehicular Technology Conference*, 2004, *VTC2004-Fall*, 2004 *IEEE 60th*, 26–29 September 2004, Vol. 6, pp. 3905–3907.
21. B. Maoz, L. R. Reynolds, A. Oki, and M. Kumar, FM-CW radar on a single GaAs/AlGaAs HBT MMIC chip, *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, June 1991, pp. 3–6.
22. W. Menzel, D. Pilz, and R. Lererer, A 77-GHz FM/CW radar front-end with a low-profile low-loss printed antenna, *IEEE Transactions on Microwave Theory and Techniques*, Vol. 47, December 1999, pp. 2237–2241.
23. F. E. Nathanson, *Radar Design Principles*, New York: McGraw-Hill, 1991, pp. 448–467.
24. P. E. Pace, *Detecting and Classifying Low Probability of Intercept Radar*, Norwood, MA: Artech House, Book and CD-ROM edition, 2003.
25. S. O. Piper, MMW seekers, in *Principles and Applications of Millimeter Wave Radar*, N. C. Currie and C. E. Brown (Eds.), Norwood, MA: Artech House, 1987, chap. 14.
26. P. Proctor, Low-level collision avoidance tested, *Aviation Week and Space Technology*, Vol. 144, February 3, 1997, p. 58.
27. Quinstar Technology, Inc. *Application Note Number 5, Millimeter Wave Radar Subsystems*, http://www.quinstar.com/an05_millimeter_wave_radar_subsystems.html
28. K. Raymer and T. Weingartner, Advanced terrain data processor, *Proceedings of the IEEE/AIAA 13th Digital Avionics Systems Conference* (*DASC*), 31 October–3 November 1994, pp. 636–639.
29. Saab, 2005, http://www.products.saab.se/PDBWeb/GetFile.aspx?pathtype=ProductFiles& filetype=Files& id=1118
30. W. K. Saunders, CW and FM radar, in *Radar Handbook*, M. I. Skolnik (Ed.), New York: McGraw-Hill, 1990.
31. G. E. Stratakos, P. Bourgas, and K. Gotsis, A low cost, high accuracy radar altimeter, *Microwave Journal*, Vol. 43, February 2000, pp. 120–128.
32. L. M. Tozzi, *Resolution in frequency-modulated radars*, Ph.D. thesis, University of Maryland, College Park, 1972.
33. F. T. Ulaby and C. Elachi, *Radar Polarimetry for Geoscience Applications*, Norwood, MA: Artech House, 1990, pp. 193–200.
34. J. D. Will, Monopulse Doppler radar for vehicle applications, *Proceedings of the Intelligent Vehicles '95 Symposium*, 25–26 September 1995, p. 42.
35. J. C. Wiltse, S. P. Schlesinger, and C. M. Johnson, Back-scattering characteristics of the sea in the region from 10 to 50 GHz, *Proceedings of the IRE*, Vol. 45, February 1957, pp. 220–228.
36. S. Yamano, H. Higashida, M. Shono, S. Matsui, T. Tamaki, H. Yagi, and H. Asanuma, 76 GHz millimeter wave automobile radar using single chip MMIC, *Fujitsu Ten Technical Journal*, No. 23, 2004, pp. 12–19, http://www.fujitsu-ten.co.jp/english/company/researchdev/gihou/en_pdf/23/23-2.pdf

Further Information

1. For a general treatment, including analysis of clutter effects, Nathanson's (1991) book is very good and generally easy to read. For extensive detail and specific numbers in various actual cases, Saunders (1990) gives good coverage. The treatment of millimeter-wave seekers by Piper (1987) is both comprehensive and easy to read.

15 Pulse Radar

Melvin L. Belcher, Jr.
Northrop Grumman Corporation

Josh T. Nessmith
Georgia Tech Research Institute

15.1 Overview of Pulsed Radars

15.1.1 Basic Concept of Pulse Radar Operation

The basic operation of a pulse radar is depicted in Figure 15.1. The radar transmits pulses superimposed on a radio frequency (RF) carrier and then receives returns (reflections) from desired and undesired scatterers. Desired scatterers corresponding to targets may include space, airborne, and sea- and/or surface-based vehicles as well as the earth's surface and the atmosphere in remote sensing applications. Returns from undesired scatterers are termed *clutter.* Clutter sources encountered by radars designed to detect and track individual targets include the earth's surface, natural and man-made discrete objects, and volumetric atmospheric phenomena such as rain and birds. Short-range/low-altitude radar operation is often constrained by clutter since the multitude of undesired returns masks returns from targets of interest such as aircraft.

The range, azimuth angle, elevation angle, and range rate can be directly measured from a return to estimate target metrics, position and velocity, to support tracking. Signature data can be extracted to support noncooperative target identification or environmental remote sensing by measuring the amplitude, phase, and polarization of the returns.

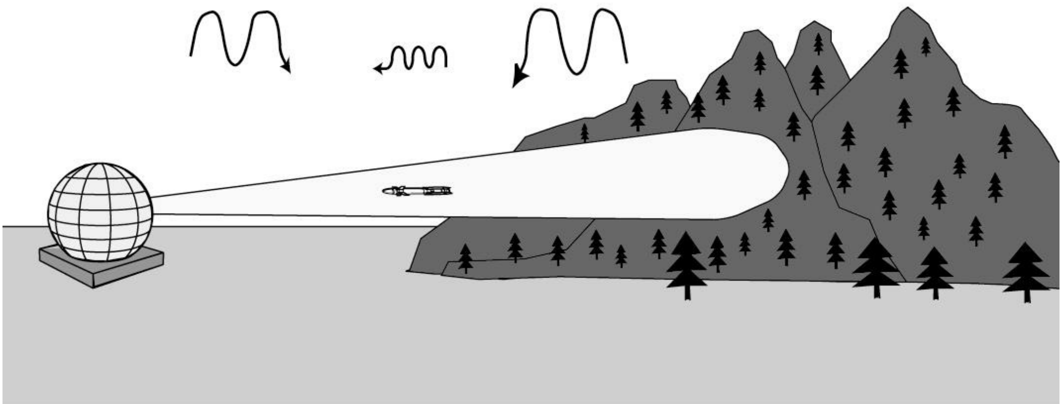


FIGURE 15.1 Pulse radar.

TABLE 15.1 Radar Bands

Band	Frequency Range	Principal Applications
HF	3–30 MHz	Over-the-horizon radar
VHF	30–300 MHz	Long-range search
UHF	300–1000 MHz	Long-range surveillance
L	1000–2000 MHz	Long-range surveillance
S	2000–4000 MHz	Surveillance
		Long-range weather characterization
		Terminal air traffic control
C	4000–8000 MHz	Fire control
		Instrumentation tracking
X	8–12 GHz	Fire control
		Air-to-air missile seeker
		Marine radar
		Airborne weather characterization
Ku	12–18 GHz	Short-range fire control
		Remote sensing
Ka	27–40 GHz	Remote sensing
		Weapon guidance
V	40–75 GHz	Remote sensing
		Weapon guidance
W	75–110 GHz	Remote sensing
		Weapon guidance

Pulse radar affords a great deal of design and operational flexibility. Pulse duration, pulse rate, and pulse bandwidth can be tailored to specific applications to provide optimal performance. Modern computer-controlled multiple-function radars exploit this capability by choosing the best *waveform* from a repertoire for a given operational mode and interference environment automatically.

15.1.2 Radar Applications

The breadth of pulse radar applications is summarized in Table 15.1 in terms of operating frequency bands. In practice, radar applications must often compete with telecommunication and other users so that the spectrum actually available for radar applications is only a portion of the depicted band. Radar applications can be grouped into search, track, and signature measurement applications. Search radars are used for target acquisition and surveillance but have relatively large range and angle tracking errors. The search functions favor broad beamwidths, low bandwidths, and low measurement update rates in order to efficiently search over a large spatial volume. As indicated in Table 15.1, search is preferably performed in the lower frequency bands. The antenna pattern is typically narrow in azimuth and has a cosecant pattern in elevation to provide acceptable coverage from the horizon to the zenith.

Tracking radars are typically characterized by a narrow beamwidth and moderate bandwidth in order to provide accurate range and angle measurements on a given target. The antenna pattern is a pencil beam with approximately the same dimensions in azimuth and elevation. Tracking is usually conducted at the higher frequency bands in order to minimize the beamwidth for a given antenna aperture area. Tracking radars are also designed to provide an adequate measurement update rate to accommodate anticipated target dynamics. A track nominally consists of estimated position and velocity based on a sequence of measurements processed by a sequential estimation algorithm termed a track filter. Track filtering smoothes the data to refine the estimate of target position and velocity. The track filter also predicts the target's flight path to provide range gating and antenna pointing control to the radar system as well as support higher echelon functions such as battle management.

Signature measurement applications include remote sensing of the environment as well as the measurement of target characteristics. In some applications, synthetic aperture radar (SAR) imaging is conducted from aircraft or satellites to characterize land usage over broad areas. Moving targets that present changing aspect to the radar can be imaged from airborne or ground-based radars via inverse synthetic aperture radar (ISAR) techniques. As defined in Section 15.3.6, cross-range resolution improves with increasing antenna extent. SAR/ISAR effectively substitutes an extended observation interval over which coherent returns are collected from different target aspect angles for a large antenna structure that would not be physically realizable in many instances.

In general, characterization performance improves with increasing frequency because of the associated improvement in range, range rate, and cross-range resolution. However, phenomenological characterization to support environmental remote sensing may require data collected across a broad swath of frequencies.

A multiple-function *phased-array* radar integrates these functions to some degree. Its design is generally driven by the track function. The degree of signature measurement implemented to support such functions as noncooperative target identification depends on the resolution capability of the radar as well as the operational user requirements. Multiple-function radar design represents a compromise among these different requirements. However, implementation constraints, multiple-target handling requirements, and reaction time requirements often dictate the use of phased-array radar systems integrating search, track, and characterization functions. The operational frequency of a multiple-function radar is a compromise between the lower frequency of the search radar and the higher frequency desired for the tracking radar. However, demanding applications may utilize a suite of low-band and high-band phased-array radar systems.

15.2 Critical Subsystem Design and Technology

The major subsystems making up a pulse radar system are depicted in Figure 15.2. The associated interaction between function and technology is summarized in this subsection.

15.2.1 Antenna

The radar antenna function is to first provide spatial directivity for the transmitted signal and then to intercept the reflected signal from a target. This is the process of transmit and receive beamforming. Most radar antennas may be categorized as mechanically scanning or electronically scanning. Mechanically scanned reflector antennas are used in applications where rapid beam scanning is not required. Electronic scanning antennas include phased arrays and frequency scanned antennas. Phased-array beams can be

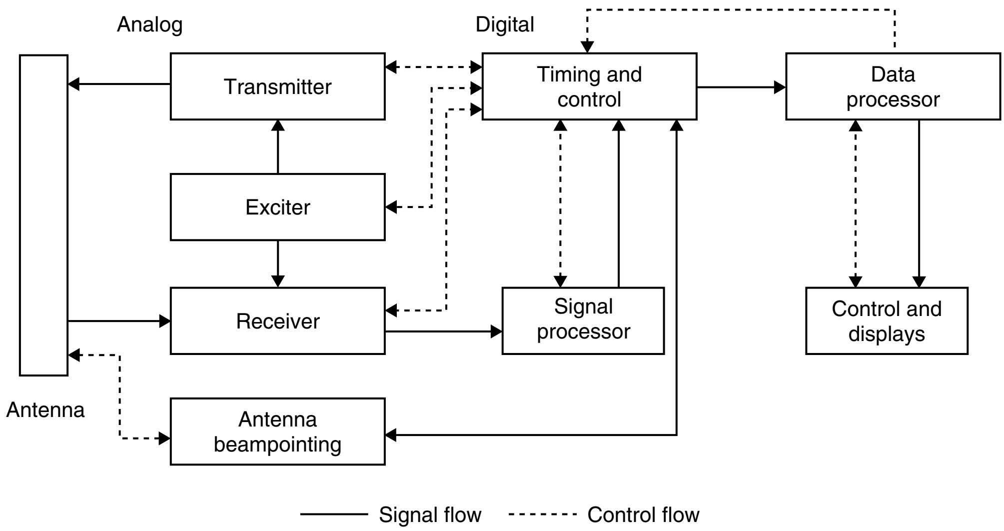


FIGURE 15.2 Radar system architecture.

steered to any point in their field-of-view, typically within 10–100 μs, depending on the latency of the beam steering subsystem and the switching time of the phase shifters. Phased arrays are desirable in multiple-function radars since they can interleave search operations with multiple target tracks.

There is a two-dimensional Fourier transform relationship between the antenna illumination function (the distribution of current across the face of the antenna) and the far-field antenna pattern. Analogous to spectral analysis, tapering the illumination function to concentrate power near the center of the antenna suppresses sidelobes while reducing the effective antenna aperture area resulting in some loss in signal-to-noise ratio (S/N) and broadening of the antenna mainlobe.

Similarly, the phase and amplitude control of the antenna illumination determines the achievable sidelobe suppression and angle measurement accuracy. Errors in the illumination due to mechanical and electrical perturbations distort the illumination function and constrain performance in these areas. Mechanical illumination error sources include antenna shape deformation due to sag and thermal effects as well as manufacturing defects. Electrical illumination error is of particular concern in phased arrays where sources include beam steering computational error and phase shifter quantization. Control of both the mechanical and electrical perturbation errors is the key to both low sidelobes and highly accurate angle measurements. Control requires that either tolerances are closely held and maintained or that there must be some means for monitoring and correction. Phased arrays are attractive for low sidelobe applications since they can provide element-level phase and amplitude control.

15.2.2 Transmitter

The transmitter function is to amplify waveforms to a power level sufficient for target detection and estimation. Tube-based transmitters are increasingly giving way to solid-state transmitters. In particular, solid-state transmit/receive modules appear attractive for constructing phased-array radar systems. In this case, each radiating element is driven by a module that contains a solid-state transmitter, phase shifter, low-noise receive amplifier, and associated control components. Active electronically scanned arrays (AESAs) built from such modules appear to offer significant reliability, mass, and implementation flexibility advantages over radar systems driven from a single transmitter. Microwave tube technology offers substantial advantages in power output but performance trends and production base considerations clearly favor solid state in new designs.

15.2.3 Receiver and Exciter

This subsystem contains the precision timing and frequency reference source or sources used to derive the master oscillator and local oscillator reference frequencies. These reference frequencies are used to down-convert received signals in a multiple-stage superheterodyne architecture to accommodate signal amplification and interference rejection. Filtering is conducted at the carrier and intermediate frequencies in processing to reject interference outside the operating band of the radar. The receiver front end is typically protected from overload during transmission through the combination of a circulator and a transmit/receive switch.

The exciter generates the waveforms for subsequent transmission. As in signal processing, the trend is toward programmable digital signal synthesis because of the associated flexibility and performance stability. Radar systems that employ Doppler filtering to detect small targets masked by large clutter returns employ extremely stable frequency reference sources in order to minimize extended clutter sidebands.

The timing and control subsystem is typically integrated with the receiver and exciter to generate and distribute signals to synchronize and command the other subsystems. This timing and control subsystem typically functions as the two-way interface between the data processor and the other radar subsystems. The increasing inclusion of BIT (built-in-test) and built-in calibration capability in timing and control subsystem designs promises to result in significant improvement in fielded system performance.

15.2.4 Signal and Data Processing

Digital processing is often divided between two processing subsystems, that is, signals and data, according to the algorithm structure and throughput demands. Signal processing includes pulse compression, Doppler filtering, and detection threshold estimation and testing. Data processing includes track filtering, user interface support, and such specialized functions as control of the electronic protection to suppress interfering signals, as well as the resource management process required to control the radar system.

The signal processor is often optimized to perform the repetitive complex multiply-and-add operations associated with the fast Fourier transform (FFT). FFT processing is used for implementing *pulse compression* via fast convolution and for Doppler filtering. Pulse compression consists of matched filtering on receive to an intrapulse modulation imposed on the transmitted pulse. As delineated subsequently, the imposed intrapulse bandwidth determines the range resolution of the pulse while the modulation format determines the suppression of the *waveform* matched-filter response outside the nominal range resolution extent. Fast convolution consists of taking the FFT of the digitized receiver output, multiplying it by the stored FFT of the desired filter function, and then taking the inverse FFT of the resulting product. Fast convolution results in significant computational saving over performing the time-domain convolution of returns with the filter function corresponding to the matched filter. The signal processor output can be characterized in terms of range gates and Doppler filters corresponding approximately to the range and Doppler resolution, respectively.

The signal processor is designed to output *detection reports* corresponding to returns that pass an adaptive criteria process intended to discriminate targets from clutter. This detection decision is typically performed on the basis of relative magnitude between a given return and the local background interference. More sophisticated algorithmic techniques are required to detect small targets embedded in large clutter returns.

The radar data processor typically consists of a general-purpose computer with a software infrastructure designed to support real-time operation. Fielded radar data processors range from microcomputers to mainframe computers, depending on application requirements. Data processor software and hardware requirements are sometimes mitigated by off-loading timing and control functions to specialized hardware. The trend is toward increasing use of commercial off-the-shelf digital processing elements for radar applications and tighter integration of the signal and data processing functions bolstered by periodic software upgrades.

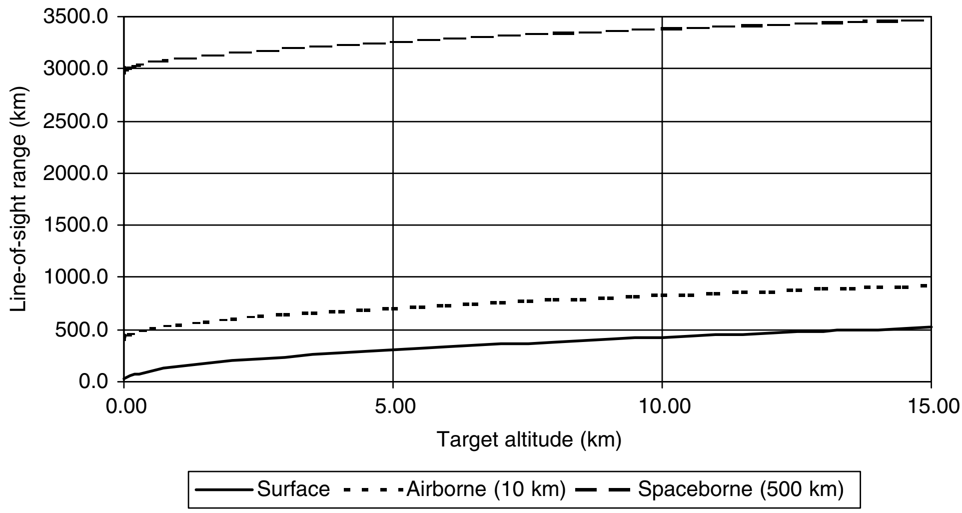


FIGURE 15.3 Maximum line-of-sight range for surface-based radar, an airborne surveillance radar, and a space-based radar.

15.3 Radar Performance Prediction

15.3.1 Radar Line-of-Sight

With the exception of over-the-horizon (OTH) radar systems, which exploit either sky-wave bounce or ground-wave propagation modes and sporadic ducting effects at higher frequencies, surface and airborne platform radar operation is limited to the refraction-constrained line of sight. Atmospheric refraction effects can be closely approximated by setting the earth's radius to 4/3 its nominal value in estimating horizon-limited range. The resulting line-of-sight range is depicted in Figure 15.3 for a surface-based radar, an airborne surveillance radar, and a space-based radar.

As evident in the plot, airborne and space-based surveillance radar systems offer significant advantages in the detection of low-altitude targets that would otherwise be masked from surface-based radars by earth curvature and terrain features. However, clutter filtering techniques must be used in order to detect targets since surface clutter returns will be present at almost all ranges of interest.

15.3.2 Radar Range Equation

The radar range equation is commonly used to estimate radar system performance, given that line-of-sight conditions are satisfied. This formulation essentially computes the S/N at the output of the radar signal processor. In turn, S/N is used to provide estimates of radar detection and position measurement performance as described in subsequent subsections. S/N can be calculated in terms of the number of pulses coherently integrated over a single coherent processing interval (CPI) presupposing matched filtering such that

$$\frac{S}{N} = \frac{PD\,A_r\sigma\,T_pN_p}{(4\pi)^2R^4\kappa\,T_sL_tL_{rn}L_{sp}} \tag{15.1}$$

where P is peak transmitter power output, D is directivity of the transmit antenna, A_r is effective aperture area of the receive antenna in meters squared, T_p is pulse duration, σ is *radar cross section* in square meters, N_p is the number of coherently integrated pulses within the coherent processing interval, R is range to target in meters, L_t is system ohmic and nonohmic transmit losses, L_{rn} is system nonohmic receive losses due to electrical mismatch among components, L_{sp} is signal processing losses such as imposed by sidelobe weighting and detection thresholding, κ is Boltzmann's constant ($\sim 1.38 \times 10^{-23}$ °K), and T_s is system noise temperature, including receive ohmic losses (Kelvin). Ohmic loss denotes losses imposed by resistive effects that contribute to system noise as well as attenuate desired signals (Blake, 1986).

At X-band and above it may also be necessary to include propagation loss due to atmospheric attenuation as an additional loss term (Blake, 1986). This form of the radar range equation is applicable to radar systems using pulse compression or pulse Doppler waveforms as well as the unmodulated single-pulse case. In many applications, average power is a better measure of system performance than peak power since it indicates the S/N improvement achievable with pulse integration over a given interval of time. Hence, the radar range equation can be modified such that

$$\frac{S}{N} = \frac{P_a D_t A_r \sigma T_c}{(4\pi)^2 R^4 \kappa T_s L_t L_{rn} L_{sp}} \tag{15.2}$$

where P_a is average transmitter power and T_c is coherent processing interval (CPI) over which pulses are coherently integrated to support Doppler filtering. Noncoherent pulse integration does not enhance Doppler resolution and is less efficient for large numbers of pulses.

The portion of time over which the transmitter is in operation is referred to as the radar duty cycle. The average transmitter power is the product of duty cycle and peak transmitter power. Duty cycle ranges from less than 1% for typical *noncoherent* pulse radars to somewhat less than 50% for high pulse repetition frequency (PRF) pulse Doppler radar systems.

15.3.3 Antenna Directivity and Aperture Area

The directivity of the antenna is

$$D = \frac{\eta 4\pi A}{\lambda^2} \tag{15.3}$$

where A is the antenna aperture area, η is aperture efficiency, and λ is radar carrier wavelength. Aperture inefficiency is due to the antenna illumination factor imposed by space fed illumination or sidelobe-reduction tapering.

The common form of the radar range equation uses power gain rather than directivity. Antenna gain is equal to the directivity divided by the antenna losses. In the design and analysis of modern radars, directivity is a more convenient measure of performance because it permits designs with distributed active elements, such as solid-state phased arrays, to be assessed to permit direct comparison with passive antenna systems. Directivity provides the power density relative to an isotropic radiator. Beamwidth and directivity are inversely related; a highly directive antenna will have a narrow beamwidth. For typical design parameters, the following approximation is often used

$$D \approx \frac{30000}{\beta_{az}\beta_{el}} \tag{15.4}$$

where β_{az} and β_{el} are the radar azimuth and elevation beamwidths, respectively, in degrees.

15.3.4 Radar Cross Section

In practice, the *radar cross section* (RCS) of a realistic target must be considered a random variable with an associated correlation interval. Targets are composed of multiple interacting scatters so that the composite return varies in magnitude with the constructive and destructive interference of the contributing returns. The target RCS is typically estimated as the mean or median of the target RCS distribution. The associated correlation interval indicates the rate at which the target RCS varies over time. RCS fluctuation degrades target detection performance at moderate to high probability of detection.

TABLE 15.2 Median Target RCS (m^2)

Carrier Frequency (GHz)	1–2	3	5	10	17
Aircraft (nose/tail avg.)					
Small propeller	2	3	2.5		
Small jet (Lear)	1	1.5	1	1.2	
T38-twin jet, F5	2	2–3	2	1–2/6	
T39-Sabreliner	2.5			10/8	9
F4, large fighter	5–8/5	4–20/10	4	4	
737, DC9, MD80	10	10	10	10	10
727, 707, DC8-type	22–40/15	40	30	30	
DC-10-type, 747	70	70	70	70	
Ryan drone			2/1		
Standing man (180 lb)	0.3	0.5	0.6	0.7	0.7
Automobiles	100	100	100	100	100
Ships-incoming ($\times 10^4$ m^2)					
4K tons	1.6	2.3	3.0	4.0	5.4
16K tons	13	18	24	32	43
Birds					
Sea birds	0.002	0.001–0.004	0.004		
Sparrow, starling, and so on	0.001	0.001	0.001	0.001	0.001

Note: Slash marks indicate different set.
Source: F.E. Nathanson, *Radar Design Principles*, 2nd ed., New York: McGraw-Hill, 1991. With permission.

The median RCS of typical targets is given in Table 15.2. The composite RCS measured by a radar system may be composed of multiple individual targets in the case of closely spaced targets such as a bird flock.

15.3.5 Loss and System Temperature Estimation

Sources of S/N loss include ohmic and nonohmic (mismatch) loss in the antenna and other radio frequency components, propagation effects, signal processing deviations from matched filter operation, detection thresholding, and search losses. Scan loss in phased-array radars is due to the combined effects of the decrease in projected antenna area and element mismatch with increasing scan angle. Search operations impose additional losses due to target position uncertainty. Because the target position is unknown before detection, the beam, range gate, and Doppler filter will not be centered on the target return. Hence, straddling loss will occur as the target effectively straddles adjacent resolution cells in range and Doppler. Beam-shape loss is a consequence of the radar beam not being pointed directly at the target so that there is a loss in both transmit and receive antenna gain. In addition, detection threshold loss associated with radar system adaptation to interference must be included (Nathanson, 1991).

System noise temperature estimation corresponds to assessing the system thermal noise floor referenced to the antenna output. Assuming the receive hardware is at ambient temperature and that initial receive gain is sufficient to establish the noise floor, the system noise temperature can be estimated as

$$T_s = T_a + 290 \times (L_r F - 1) \tag{15.5}$$

where T_a is the antenna noise temperature, L_r is receive ohmic losses, and F is the receive amplifier noise figure. A more complex analysis is often required for active electronically scanned arrays where external noise sources must be considered as well as beamformer noise contributions and multiple stages of amplification within the receive chain.

Table 15.3 provides representative loss and noise temperature budgets for several major radar classes. In general, loss increases with the complexity of the radar hardware between the transmitter/receiver and the antenna radiator. Reflector antennas and active phased arrays impose relatively low loss, while passive

TABLE 15.3 Typical Microwave Loss and System Temperature Budgets

	Mechanically Scanned Reflector Antenna	Electronically Scanned Slotted Array	Solid-State Phased Array
Nominal losses			
Transmit loss, L_t (dB)	1	1.5	0.5
Nonohmic receiver loss, L_r (dB)	0.5	1.0	0.5
Signal processing loss, L_{sp} (dB)	1.4	1.4	1.4
Scan loss (dB)	N/A	−30 log [cos (scan angle)]	−30 log [cos (scan angle)]
Search losses, L_{DS}			
Beam shape (dB)	3	3	3
Range gate straddle (dB)	0.5	0.5	0.5
Doppler filter straddle (dB)	0.5	0.5	0.5
Detection thresholding (dB)	1	1	1
System noise temperature (in °K)	500	1000	500

TABLE 15.4 Resolution and Accuracy

Dimension	Nominal Resolution	Noise-Limited Accuracy
Angle	$\frac{\alpha\lambda}{d}$	$\frac{\alpha\lambda}{dK_m\sqrt{2S/N}}$
Range	$\frac{\alpha C}{2B}$	$\frac{\alpha C}{2BK_i\sqrt{2S/N}}$
Doppler	$\frac{\alpha}{\text{CPI}}$	$\frac{\alpha}{\text{CPI}K_i\sqrt{2S/N}}$
SAR/ISAR	$\frac{\alpha\lambda}{2\cdot\Delta\theta}$	$\frac{\alpha\lambda}{2\cdot\Delta\theta K_i\sqrt{2S/N}}$

Note: α taper broadening factor, typically ranging from 0.89 (unweighted) to 1.3 (Hamming); d, antenna extent in azimuth/elevation; B, *waveform* bandwidth; K_m, monopulse slope factor, typically on the order of 1.5; K_i, interpolation factor, typically on the order of 1.8; θ, line-of-sight rotation of target relative to radar over CPI.

array antennas impose relatively high loss. The monopulse difference channels may possess higher ohmic loss than the sum channel mandating a corresponding analysis of system noise temperature in order to accurately assess angle-of-arrival measurement precision.

15.3.6 Resolution and Accuracy

The fundamental resolution capabilities of a radar system are summarized in Table 15.4. In general, there is a trade-off between mainlobe resolution corresponding to the nominal range, Doppler, and angle resolution, and effective dynamic range corresponding to suppression of sidelobe components. This is evident in the use of weighting to suppress Doppler sidebands and angle sidelobes at the expense of broadening the mainlobe and S/N loss.

Cross range denotes either of the two dimensions orthogonal to the radar line-of-sight (RLOS). Cross-range resolution in real-aperture antenna systems is closely approximated by the product of target range and radar beamwidth in radians. Attainment of the nominal ISAR/SAR cross-range resolution requires coherent signal processing over a sequence of returns corresponding to a rotation of the target orientation relative to the RLOS or as in SAR with the linear translation of the antenna aboard aircraft or satellites.

Nonlinear or adaptive algorithmic techniques may be required to generate a focused image in order to compensate for imaging artifacts such as scatterer change in range over the CPI.

The best accuracy performance occurs for the case of thermal noise-limited error. The resulting accuracy is proportional to the resolution of the radar divided by the square root of the S/N. In this formulation, the single-pulse S/N has been multiplied by the number of pulses integrated within the CPI as indicated in Equations 15.1 and 15.2. In practice, accuracy is constrained by environmental effects, target characteristics, and instrumentation error effects that may dominate noise-limited. Environmental effects include multipath and refraction. Target glint is characterized by an apparent wandering of the target position because of coherent interference effects associated with the composite return from the individual scattering centers on the target. Instrumentation error due to channel imbalance and similar effects is minimized with alignment and calibration but may constrain the ultimate tracking accuracy.

15.3.7 Radar Range Equation for Search and Track

The radar range equation can be modified to directly address performance in the two primary radar missions: search and track.

Search performance is basically determined by the capability of the radar system to detect a target of specific RCS at a given maximum detection range while scanning a given solid angle extent within a specified period of time. S/N can be set equal to the minimum value required for a given detection performance, $(S/N)_{\text{req}}$ while R can be set to the maximum required target detection range, R_{max}. Manipulation of the radar range equation results in the following expression:

$$\frac{P_{\text{a}}A}{L_{\text{t}}L_{\text{rn}}L_{\text{sp}}L_{\text{os}}T_{\text{s}}} \geq \left(\frac{S}{N}\right)_{\text{req}} \frac{R_{\text{max}}^4 \Omega}{\sigma T_{\text{fs}}} 16\kappa \tag{15.6}$$

where Ω is the solid angle over which search must be performed (steradians), T_{fs} is the time allowed to search Ω by operational requirements, and L_{os} is the composite incremental loss associated with search.

The left-hand side of the equation contains radar design parameters, while the right-hand side is determined by target characteristics and operational requirements. The right-hand side of the equation is evaluated to determine radar requirements. The left-hand side of the equation is evaluated to determine if the radar design meets the requirements.

The track radar range equation is conditioned on noise-limited angle accuracy as this measure stresses radar capabilities significantly more than range accuracy in almost all cases of interest. The operational requirement is to maintain a given data rate track providing a specified single-measurement angle accuracy for a given number of targets with specified RCS and range. Antenna beamwidth, which is proportional to the radar carrier wavelength divided by antenna extent, impacts track performance since the degree of S/N required for a given measurement accuracy decreases as the beamwidth decreases. Track performance requirements can be bounded as

$$\frac{P_{\text{a}}A^3}{\lambda^4 L_{\text{t}}L_{\text{rn}}L_{\text{sp}}L_{\text{os}}T_{\text{s}}}\kappa_{\text{m}}^2\eta^2 \geq \frac{R_{\text{max}}^4 r N_{\text{t}}}{\sigma\sigma_{\phi}^2} 5\kappa \tag{15.7}$$

where r is the single-target track rate, N_{t} is the number of targets under track in different beams, and σ_{ϕ} is the required angle accuracy standard deviation (radians). In general, a phased-array radar antenna is required to support multiple target tracking when $N_{\text{t}} > 1$.

Incremental search losses are suppressed during single-target-per-beam tracking. The beam is pointed as closely as possible to the target to suppress beamshape loss. The tracking loop centers the range gate and Doppler filter on the return. Detection thresholding loss can be minimal since the track range window is small though the presence of multiple targets motivates continual detection processing. The antenna receive loss in a phased-array radar may be somewhat higher for the track radar range equation since the weighting function to achieve sidelobes to meet angular accuracy for the track function are often higher than for the search function

15.4 Radar Waveforms

15.4.1 Pulse Compression

Typical pulse radar waveforms are summarized in Table 15.5. In most cases, the signal processor is designed to closely approximate a matched filter. As indicated in Table 15.4, the range and Doppler resolution of any match-filtered *waveform* are inversely proportional to the *waveform* bandwidth and duration, respectively. Pulse compression, using modulated waveforms, is attractive since S/N is proportional to pulse duration rather than bandwidth in matched filter implementations. Ideally, the intrapulse modulation is chosen to attain adequate range resolution and range sidelobe suppression performance while the pulse duration is chosen to provide the required sensitivity. Pulse compression waveforms are characterized as having a time bandwidth product (TBP) significantly greater than unity, in contrast to an unmodulated pulse, which has a TBP of approximately unity.

15.4.2 Pulse Repetition Frequency

The radar system pulse repetition frequency (PRF) determines its ability to unambiguously measure target range and range rate in a single CPI as well as determining the inherent clutter rejection capabilities of the radar system. In order to obtain an unambiguous measurement of target range, the interval between radar pulses (1/PRF) must be greater than the time required for a single pulse to propagate to a target at a given range and back. The maximum unambiguous range is then given by $C/(2 \cdot \text{PRF})$ where C is the velocity of electromagnetic propagation.

Returns from moving targets and clutter sources are offset from the radar carrier frequency by the associated Doppler frequency. As a function of range rate, R', the Doppler frequency, f_d, is given by $2R'/\lambda$. A coherent pulse train complex samples the returns' Doppler modulation at the PRF. Returns are folded in frequency if the PRF is less than the target Doppler.

Clutter returns are primarily from stationary or near-stationary surfaces such as terrain. In contrast, targets of interest often have a significant range rate relative to the radar clutter. Doppler filtering can suppress returns from clutter. With the exception of frequency ambiguity, the Doppler filtering techniques used to implement pulse Doppler filtering are quite similar to those used in continuous wave radar applications. Ambiguous measurements can be resolved over multiple CPIs by using a sequence of slightly different PRFs and correlating detections among the CPIs.

15.4.3 Detection and Search

Detection processing generally consists of comparing the amplitude of each range gate/Doppler filter output with a threshold. A detection is reported if the amplitude exceeds that threshold. A false alarm occurs when noise or other interference produces an output of sufficient magnitude to exceed the detection threshold. As the detection threshold is decreased, both the detection probability and the false alarm probability increase. S/N must be increased to enhance detection probability while maintaining a constant false alarm probability.

RCS fluctuation effects due to the coherent summation of returns from unresolved scatterer must be considered in assessing detection performance. The Swerling models which use chi-square probability density functions (PDFs) of 2 and 4 degrees of freedom (DOF) are commonly used for this purpose. The Swerling 1 and 2 models are based on the 2 DOF PDF and can be derived by modeling the target as an ensemble of independent scatterers of comparable magnitude. This model is considered representative of complex targets such as aircraft. The Swerling 3 and 4 models use the 4 DOF PDF and correspond to a target with a single dominant scatterer and an ensemble of lesser scatterers. The Swerling 1 and 3

TABLE 15.5 Selected *waveform* Characteristics

	Comments	Time Bandwidth Product	Range Sidelobes (dB)	S/N Loss (dB)	Range/Doppler Coupling	ECM/EMI Robustness
Unmodulated	No pulse compression	~ 1	Not applicable	0	No	Poor
Linear frequency modulation	Linearly swept over band width	>10	Unweighted: -13.5 Weighted: >-40[a]	0 0.7–1.4	Yes	Poor
Nonlinear frequency modulation	Multiple variants	*waveform* specific	*waveform* specific	0	*waveform* specific	Fair
Barker	N-bit biphase	$\leq 13(N)$	$-20\log(N)$	0	No	Fair
Linear recursive sequence	N-bit biphase N-bit polyphase ($N = \text{integer}^2$)	$-N$; $\sim N$	$\sim -10\log(N)$ $\sim -10\log(p^{2N})$	0 0	No Limited	Good Good
Frequency coding	N subpulses noncoincidental in time and frequency	$\sim N^2$	*waveform* specific, periodic and pseudorandom	0.7–1.40 0	*waveform* specific	Good

[a] Constraint due to typical technology limitations rather than fundamental *waveform* characteristics.

models presuppose slow fluctuation such that the target RCS is constant from pulse to pulse within a scan. In contrast, the RCS of Swerling 2 and 4 targets is modeled as independent on a pulse to pulse basis. The selection of fast or slow fluctuation models is driven by the radar measurement update and scan rates relative to the effective decorrelation interval of the target. Frequency agility may be used to transform a target from a slow to fast fluctuation behavior when it is desired to increase the number of independent observations in a given time interval to facilitate detection. The Swerling 1/2 model is often used to establish a conservative performance estimate as well as to facilitate analytical convenience. However, the Swerling 3/4 model is often more representative of aerodynamic targets viewed from their front quarter.

Single-pulse detection probabilities for nonfluctuating, Swerling 1/2, and Swerling 3/4 targets are depicted in Figure 15.4. This curve is based on a typical false alarm number corresponding approximately to a false alarm probability of 10^{-6}. The difference in S/N required for a given detection probability for a fluctuating target relative to the nonfluctuating case is termed the fluctuation loss.

The detection curves presented here and in most other references presuppose noise-limited operation. In many cases, the composite interference present at the radar system output will be dominated by clutter returns or electromagnetic interference such as that imposed by hostile electronic countermeasures. The standard textbook detection curves cannot be applied in these situations unless the composite interference

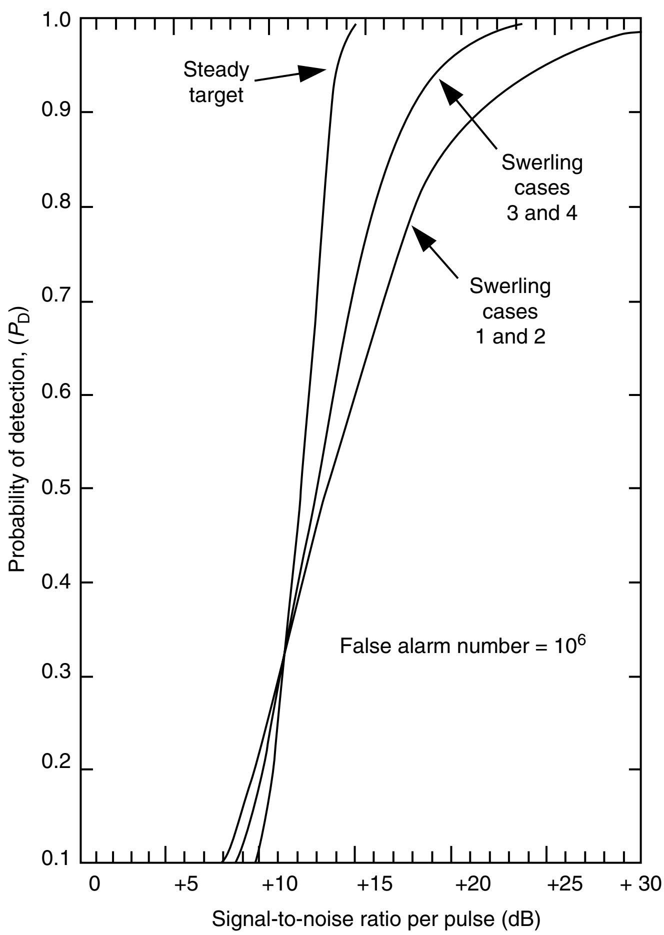


FIGURE 15.4 Detection probabilities for various target fluctuation models. (From F.E. Nathanson, *Radar Design Principles*, 2nd ed., New York: McGraw-Hill, 1991, p. 91. With permission.)

is statistically similar to thermal noise with a Gaussian PDF and a white power spectral density. The presence of non-Gaussian interference increases the false alarm probability beyond the design value. Adaptive detection threshold estimation techniques are often required to search for targets in environments characterized by such interference (Nathanson, 1991). Although the Swerling RCS models are quite often used in a preliminary design of a radar system, high-fidelity models of anticipated target signatures will be used for detailed evaluation.

15.5 Estimation and Tracking

15.5.1 Measurement Error Sources

Radars measure target range and angle position, and potentially, Doppler frequency. Angle measurement performance is emphasized here since the corresponding cross-range error dominates range error for most practical applications. Target returns are smoothed in a tracking filter, but the achievable error reduction is ultimately determined by the measurement accuracy and radar system error and target maneuverability characteristics. Radar measurement error can be characterized as indicated in Table 15.6.

The radar design and the alignment and calibration process development must consider the characteristics and interaction of these error components. Integration of automated techniques to support alignment and calibration is an area of strong effort in modern radar design that can lead to significant performance improvement in fielded systems.

As indicated previously, angle measurement generally is the limiting factor in measurement accuracy. Target azimuth and elevation position is primarily measured by a monopulse technique in modern radars though early systems used sequential lobing and conical scanning. Specialized monopulse-tracking radars utilizing reflectors have achieved instrumentation and S/N angle residual systematic error as low as 50 μrad. Phased-array antennas have achieved a random error of less than 60 μrad, but the composite systematic residual errors remain to be measured. The limitations are primarily in the tolerance on the phase and amplitude of the antenna illumination function.

Figure 15.5 shows the monopulse beam patterns. The first is the received sum pattern that is generated by a feed that provides the energy from the reflector or phased-array antenna through two ports in equal amounts and summed in phase in a monopulse comparator shown in Figure 15.6. The second is the difference pattern generated by providing the energy through the same two ports in equal amounts but taken out with a phase difference of p radians, giving a null at the center. A target located at the center of the same beam would receive a strong signal from the sum pattern with which the target could be detected and ranged. The received difference pattern would produce a null return, indicating the target was at the center of the beam. If the target were off the null, the signal output or difference voltage would be almost linear and proportional to the distance off the center (off-axis), as shown in the figure. This output of the monopulse processor is the real part of the dot product of the complex sums and the difference signals

TABLE 15.6 Radar Measurement Error

Random errors	Those errors that cannot be predicted except on a statistical basis. The magnitude of the random error can be termed the *precision* and is an indication of the repeatability of a measurement
Bias errors	A systematic error whether due to instrumentation or propagation conditions. A nonzero mean value of a random error
Systematic error	An error whose quantity can be measured and reduced by calibration
Residual systematic	Those errors remaining after measurement and calibration.
Accuracy	The magnitude of the rms value of the residual systematic and random errors

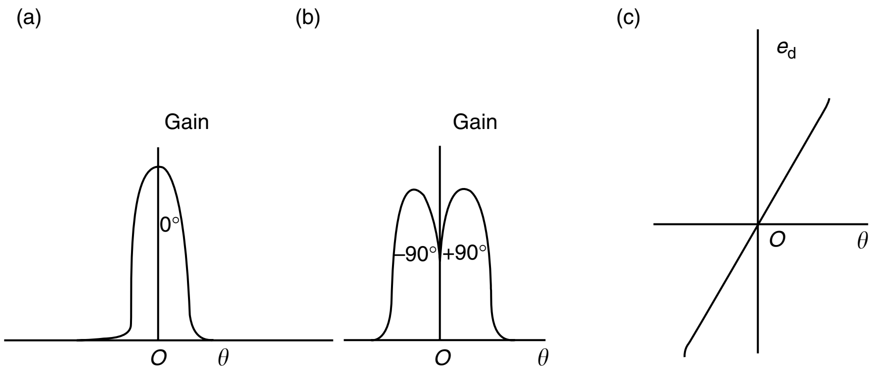


FIGURE 15.5 Monopulse beam patterns and difference voltage: (a) sum (S), (b) difference (D), and (c) difference voltage.

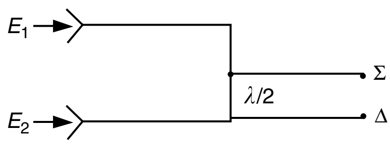


FIGURE 15.6 Monopulse comparator.

divided by the absolute magnitude of the sum signal squared, that is,

$$e_{\mathrm{d}} = \mathrm{Re}\left[\frac{\Sigma \cdot \Delta}{|\Sigma|^2}\right] \tag{15.8}$$

The random instrumentation measurement errors in the angle estimator are caused by phase and amplitude errors of the antenna illumination function. In reflector systems, such errors occur because of the position of the feedhorn, differences in electrical length between the feed and the monopulse comparator, mechanical precision of the reflector, and its mechanical rotation. In phased-array radars, these errors are a function of the phase shifters, time delay units, and combiners between the antenna elements and the monopulse comparator as well as the precision of the array. Although these errors are random, they may have correlation intervals considerably longer than the white noise considered in the thermal-noise random error and may depend on the flight path of the target. For a target headed radially from or toward the radar, the correlation period of angle-measurement instrumental errors is essentially the tracking period. For crossing targets, the correlation interval may be pulse to pulse.

As in the estimate of range, the propagation effects of refraction and multipath also enter into the tracking error. The bias error in range and elevation angle by refraction can be estimated, respectively, as

$$\begin{aligned} E_{\mathrm{R}} &= 0.007\, N_{\mathrm{s}} \text{ cosecant } E_0 \quad \text{(meters)} \\ E_{\mathrm{E}} &= N_{\mathrm{s}} \cot E_0 \quad (\mu\text{rad}) \end{aligned} \tag{15.9}$$

where N_{s} is the surface refractivity and E_0 is the true elevation angle (Barton and Ward, 1984).

One can calculate the average error in multipath. However, one cannot correct for it as in refraction since the magnitude and direction of the error cannot be accurately estimated in advance unless there are controlled conditions such as in a carefully controlled experiment. Hence, the general approach is to design the antenna sidelobes to be as low as feasible and accept the multipath error that occurs when tracking close to the horizon. There has been considerable research to find means to reduce the impact, including using very wide bandwidths to separate the direct path from the multipath return as well as specialized track filtering techniques that accommodate multipath effects.

15.6 Tracking Filter Performance

Target tracking based on processing returns from multiple CPIs can provide a target position and velocity estimate of greater accuracy than the single-CPI measurement accuracy delineated in Table 15.4. In principle, the error variance of the estimated target position with the target moving at a constant velocity is approximately $4\sigma_{\mathrm{m}}^2/n$ where n is the number of independent measurements processed by the track filter and σ_{m} is the standard deviation of the single-measurement error. In practice, the variance reduction factor afforded by a track filter is often limited to about an order of magnitude because of the reasons summarized in the following paragraphs.

Track filtering generally provides smoothing and prediction of target position and velocity via a recursive prediction-correction process. The filter predicts the target's position at the time of the next measurement based on the current smoothed estimates of position, velocity, and possibly acceleration. The subsequent difference between the measured position at this time and the predicted position is used to update the smoothed estimates. The update process incorporates a weighting vector that determines the relative significance given the track filter prediction versus the new measurement in updating the smoothed estimate.

Target model fidelity and adaptivity are fundamental issues in track filter mechanization. The performance of one-dimensional polynomial algorithms, such as the alpha–beta filter, to track targets from one pulse to the next and provide modest smoothing is adequate to maintain track. However, fixed-coefficient polynomial track filtering ignores knowledge of the equations of motion governing the target so that their smoothing and long-term prediction performance is relatively poor. In addition, simple polynomial tracking filters do not incorporate any adaptivity or measure of estimation quality.

Kalman filtering addresses these shortcomings at the cost of significantly greater computational complexity. Target equations of motion are modeled explicitly such that the position, velocity, and potentially higher-order derivatives of each measurement dimension are estimated by the track filter as a state vector. The error associated with the estimated state vector is modeled via a covariance matrix that is also updated with each iteration of the track filter. The covariance matrix determines the weight vector used to update the smoothed state vector in order to incorporate such factors as measurement S/N and dynamic target maneuvering.

Smoothing performance is constrained by the degree of a priori knowledge of the targets kinematic motion characteristics. For example, Kalman filtering can achieve significantly better error reduction against ballistic or orbital targets than against maneuvering aircraft. In the former case the equations of motion are explicitly known, while the latter case imposes motion model error because of the presence of unpredictable pilot or guidance system commands. Similar considerations apply to the fidelity of the track filters model of radar measurement error. Failure to consider the impact of correlated measurement errors may result in underestimating track error when designing the system. Tracking techniques have been developed to provide more robust performance in the presence of imperfect knowledge of measurement and maneuver characteristics.

Many modern tracking problems are driven by the presence of multiple targets and low signal-to-clutter ratio conditions which impose a need for assigning measurements to specific tracks as well as accommodating unresolved returns from closely spaced targets and clutter. Existing radars often employ some variant of the nearest-neighbor algorithm where a measurement is uniquely assigned to the track with a predicted position minimizing the normalized track filter update error. More sophisticated techniques assign measurements to multiple tracks if they cannot clearly be resolved or make the assignment on the basis on several contiguous update measurements.

Defining Terms

Coherent: Integration where magnitude and phase of received signals are preserved in summation.

Noncoherent: Integration where only the magnitude of received signals is summed.

Phased array: Antenna composed of an aperture of individual radiating elements. Beam scanning is implemented by imposing a phase taper across the aperture to collimate signals received from a given angle of arrival.

Pulse compression: The matched-filter processing of a pulsed-signal waveform with arbitrary duration and modulated so that the resulting range-delay resolution is the reciprocal of the imposed bandwidth.

Radar cross section (RCS): Measure of the reflective strength of a radar target; usually represented by the symbol s, measured in square meters, and defined as $4p$ times the ratio of the power per unit solid angle scattered in a specified direction of the power unit area in a plane wave incident on the scatterer from a specified direction.

Waveform: Transmitted signal consisting of pulsed carrier with phase modulation or frequency modulation imposed if desired to support pulse compression.

References

D.K. Barton and H.R. Ward, *Handbook of Radar Measurement,* Dedham, MA: Artech, 1984.
L.V. Blake, *Radar Range-Performance Analysis*, Dedham, MA: Artech, 1986.
G.V. Morris, *Airborne Pulsed Doppler Radar*, Dedham, MA: Artech, 1988.
F.E. Nathanson, *Radar Design Principles*, 2nd ed., New York: McGraw-Hill, 1991.

Further Information

1. M.I. Skolnik, Ed., *Radar Handbook*, 2nd ed., New York: McGraw-Hill, 1990.
2. *IEEE Standard Radar Definitions*, *IEEE Standard 686-1990*, April 20, 1990.

16

Automotive Radar

Madhu S. Gupta

San Diego State University

Scope: An automotive radar, as the name suggests, is any radar that has an application in automobiles and other autonomous ground vehicles. As a result, it represents a large and heterogeneous class of radars that are based on different technologies (e.g., laser, ultrasonic, microwave), perform different functions (e.g., obstacle and curb detection, collision anticipation, adaptive cruise control), and employ different operating principles (e.g., pulse radar, FMCW radar, microwave impulse radar). This article is limited to microwave radars that form a commercially significant subset of automotive radars. Microwave radars have an advantage over the laser and infrared radars (or "lidar") in that they are less affected by the presence of precipitation or smoke in the atmosphere. They also have the advantage of being unaffected by air temperature changes that would degrade an ultrasonic radar.

Need: The need for automotive radars can be understood at three different levels. At the national level, the statistics on traffic fatalities, injuries, and property loss due to vehicle accidents, and estimates of their fractions that are preventable with technological aids, has encouraged the development of automotive radar. The economic value of those losses, when compared with the dropping cost of automotive radar, leads to a cost-benefit analysis that favors their widespread deployment. At the level of the automotive manufacturer, radar is another "feature" for the consumer to purchase, that could be a possible source of revenue and competitive advantage. It is also a possible response to regulatory and public demands for safer vehicles. At the level of vehicle owners, automotive radar has an appeal as a safety device, and as a convenient, affordable gadget. Of greater practical importance is the potential for radar to lower the stress in driving and decrease the sensory workload of the driver by taking over some of the tasks requiring attentiveness, judgment, and skill.

Antecedents: Lower-frequency electronic systems have had a presence for decades in various automobile applications, such as entertainment, fuel injection, engine control, onboard diagnostics, antitheft systems, antiskid brakes, cruise control, and suspension control. The earliest microwave frequency products introduced in automobiles were speed radar detectors in the 1970s, followed by cellular telephones in the 1980s, and direct satellite receivers and GPS navigational aids in the 1990s. The use of microwave vehicular radars can also be traced back to the 1970s, when rear obstacle detection systems were first installed, mostly in trucks. Some of those radars evolved from simple motion detectors and security alarms that employed Gunn diodes and operated typically in the X-band around 10 GHz. The modern automotive radar owes its origin to three more recent technological advancements: low-cost monolithic microwave devices and hybrid or monolithic circuits; microprocessors, and digital signal processing.

Status: Automobile radars have been developed by a number of different companies in Europe, Japan, and the United States, since the 1970s. Automotive radars for speed measurement and obstacle detection have been available for more than a decade as an aftermarket product, and have been installed primarily on trucks and buses in quantities of thousands. Only now (in 2000) are automotive radars becoming available as original equipment for installation on trucks, buses, and automobiles. However, other types of radars are still in the development, prototyping, field-testing, and early introduction stages. Some of the most important challenges for the future in the development of automotive radar include (a) development of more advanced human interface, (b) reduction of cost, (c) meeting user expectations, and (d) understanding the legal ramifications of equipment failure.

16.1 Classification

An automotive radar can be designed to serve a number of different functions in an automobile, and can be classified on that basis as follows:

1. *Speed Measuring Radar.* Vehicle speed is measured with a variety of types of speedometers, many of which are based on measuring the rate of revolution of the wheels, and are therefore affected by tire size and wheel slippage. By contrast, a speed measuring radar can determine the true ground speed of a vehicle, that may be required for the instrument panel, vehicle control (e.g., detection of skidding and antilock braking), and other speed-dependent functions.
2. *Obstacle Detection Radar.* Such a radar is essentially a vision aid for the driver, intended to prevent accidents by monitoring regions of poor or no visibility, and providing a warning to the driver. It can perform functions such as curb detection during parking, detection of obstacles in the rear of a vehicle while backing, and sensing the presence of other vehicles in blind zones (regions around the vehicle that are not visible to the driver via side- and rear-view mirrors) during lane changing.
3. *Adaptive Cruise Control Radar.* A conventional cruise control maintains a set vehicle speed regardless of the vehicular environment, and as a result, is ill suited in a heavy traffic environment where vehicles are constantly entering and leaving the driving lane. An adaptive cruise control, as the name suggests, adapts to the vehicular environment, and maintains a headway (clearance between a vehicle and the next vehicle ahead of it) that is safe for the given speeds of both vehicles.
4. *Collision Anticipation Radar.* This class of radars senses the presence of hazardous obstacles (other vehicles, pedestrians, animals) in the anticipated path of the vehicle that are likely to cause collision, given the current direction and speed of the vehicle. It is therefore potentially useful both under conditions of poor atmospheric visibility (e.g., in fog, sandstorm), as well as poor judgment on the part of the driver (unsafe headway, high speed). Its purpose may be to warn the driver, to deploy airbags or other passive restraints, or take over the control of vehicle speed.
5. *Other Vehicular Monitoring and Control Radars.* Many other vehicular control functions, such as vehicle identification, location, fleet monitoring, station keeping, guidance, and path selection, can be performed with the aid of radar. Such radars may be placed on board the vehicle, or on the ground with the vehicle carrying a radar beacon or reflector, that may be coded to allow vehicle identification. Still other variations of automotive radars can be envisaged for special purposes, such as control of vehicles constrained to move along tracks, or joyride vehicles in an amusement park.

The following is a discussion of each of the first four types of radars, including their purpose, principle of operation, requirements imposed on the radar system and its constituent parts, limitations, and expected future developments. The emphasis is on the RF front-end of the radar, consisting of the transmitter, the receiver, and the antenna. The adaptive cruise control and the collision anticipation radars share a number of characteristics and needs; they are therefore discussed together as forward-looking radars.

16.2 History of Automotive Radar Development

The idea of automotive radars is almost as old as the microwave radars themselves. The history of their development may be divided into four phases as follows:

1. *Conceptual Feasibility Phase.* The earliest known experiments, carried out in the late 1940s soon after the Second World War, involved a wartime radar mounted on top of an automobile, with the cost of the radar far exceeding that of the automobile. The idea that the radar could be used to control a vehicle was considered sufficiently novel that the U.S. Patent Office issued patents on that basis. An example of such early efforts is a patent (# 2,804,160), entitled "Automatic Vehicle Control System," issued to an inventor from Detroit, Michigan, that was filed in January 1954 and issued in August 1957.
2. *Solid-State Phase.* The presence of a microwave tube in the radar for generating the transmitted signal, and the high voltage supply they required, was known to be a principal bottleneck that impacted the size, cost, reliability, and safety of automotive radar. Therefore, the discovery of solid-state microwave sources in the form of two-terminal semiconductor devices (transferred-electron devices and avalanche transit-time diodes) in the mid-1960s led to a resurgence of interest in developing automotive radar. Since solid-state microwave devices were then one of the more expensive components of radar, the possibility of using the device, employed simultaneously as the source of the microwave signal and as a mixer by taking advantage of its nonlinearity, was considered another desirable feature. Such radars were the subject of numerous experimental and theoretical studies, and even reached the stage of limited commercial production. Radars based on BARITT diodes were deployed on trucks in the 1970s to warn the driver of obstacles in the rear.
3. *MMIC Phase.* The production of monolithic microwave integrated circuits in the 1980s made both the microwave devices, and the associated circuitry, sufficiently small and inexpensive that the weight and cost of microwave components in radar was no longer an issue. The prototype radars in this period were based both on Gunn diodes and on GaAs MMICs that employed MESFETs as the active device and source of microwaves. At the same time, the technology of microstrip patch antennas, that are particularly convenient for automotive use, became sufficiently developed. A variety of volume production problems had to be solved in this phase, including those of maintaining frequency stability and packaging, while keeping the cost down. The focus of development thus shifted to other aspects of the radar design problem: reliability of the hardware and its output data; operability of the radar under unfavorable conditions; affordability of the entire radar unit; and extraction of useful information from radar return by signal processing. Prototype units of various types were field tested, and deployed on a limited scale both for evaluation and feedback as well as in special markets such as in bus fleets.
4. *Product Development Phase.* Automotive radar became economically viable as a potential consumer product due to the decreasing ratio of the cost of the radar to the cost of the automobile. Moreover, the economic risk of radar development decreased due to the availability of specific frequency bands for this application, and the confidence resulting from the results of some field tests. As a result, automotive radar is now (in year 2000) available as an "option" on some vehicles, and work continues on enhancing its capabilities for other applications, in human factors engineering, and for integration into the vehicle.

16.3 Speed-Measuring Radar

Operating Principle: The simplest method of measuring the true ground speed of a vehicle is by the use of a Doppler radar. In this radar, a microwave signal at frequency f is transmitted from on board the vehicle, aimed toward the ground; the signal scattered back by the ground is collected by a receiver also on board the vehicle. Given the large speed c of microwave signals, and the short distance between the antenna (typically mounted under the carriage) and the ground, the signal transit time is short. The returned signal is shifted in frequency due to Doppler effect, provided the vehicle and the scattering patch of the ground are moving with respect to each other, with a velocity component along the direction of propagation of the microwave signal. To ensure that, the microwave beam is transmitted obliquely toward the ground, making an angle q with respect to the horizontal, and in the plane formed by the ground velocity v of the vehicle and the perpendicular from the vehicle to the ground. Then the vehicle velocity component along the direction of propagation of the signal is $v \cos q$, and the Doppler frequency shift is $2f(v/c)\cos q$, proportional both to the transmitted signal frequency f and to the vehicle velocity v. The Doppler shift frequency is extracted by mixing the returned signal with the transmitted signal, and carrying out filtering and signal processing. Typically (for a carrier frequency of 24 GHz and a tilt angle $q = 30°$), it lies in the range of 35 Hz–3.5 kHz for vehicle speeds in the range of 1–100 miles per hour.

Error Sources: Several sources of error in speed estimation can be identified from the above discussion:

1. *Vehicle tilt.* Uneven loading of a vehicle, air pressure in the tires, and non-level ground can all contribute to a change in the tilt angle q of the beam, and hence in the estimated vehicle speed. A well-known technique for correcting this error is to employ a so-called Janus configuration in the forward and reverse directions (named after a Greek God with two heads). In this scheme, two microwave beams are transmitted from the transmitter, one in the forward and the other in the reverse direction, each making an angle q with the horizontal. The Doppler frequency shift has the same magnitude for each signal, but a tilt of the vehicle makes q one beam larger while simultaneously making the other smaller. The correct ground velocity, as well as the tilt angle of the vehicle, can be deduced from the sum and difference of the two Doppler shift frequencies.
2. *Nonzero beam width.* Any reasonably sized antenna will produce a beam of finite width, so that the angle q between the transmitted signal and the horizontal is not a constant. A spread in the values of q produces a corresponding spread in the values of the Doppler shift frequency.
3. *Vertical vehicle velocity.* Vibrations of the vehicle and uneven ground will cause the vehicle of have a velocity with respect to the ground in the vertical direction. This velocity also has a component in the direction of propagation of the microwave signal, and thus modulates the Doppler shift frequency.
4. *Surface roughness.* Although surface roughness is essential for producing scattering in the direction of the receiver, it introduces errors because the surface variations appear as variations of q as well as of vehicle height, and therefore an apparent vertical vehicle velocity.

Extensive signal processing is required to extract an accurate estimate of vehicle velocity in the presence of these error sources. Current Doppler velocity sensors employ digital signal processing and are capable of determining the velocity with 99% certainty.

16.4 Obstacle-Detection Radar

Purpose: A driver is unable to view two principal areas around the vehicle that are a potential source of accidents: one is behind the vehicle, and the other on the two sides immediately behind the driver. The need to view these areas arises only under specific conditions: when driving in reverse and when changing lanes. The exact boundaries of these areas depend on the style of vehicle, the placement and setting of the viewing mirrors, and the height and posture of the driver.

Mission Requirements: Since obstacle detection radar needs to operate over a small range (typically <10 m), cover a wide area, and does not need to determine the exact location of the obstacle in that area, its operating frequencies can be lower, where the antenna beamwidth is wide.

16.5 Adaptive Cruise Control Radar

Purpose: The adaptive cruise control (to be abbreviated hereafter as ACC) radar is so called because it not only controls the speed of a vehicle but also adapts to the speed of a vehicle ahead. The ACC radar controls the vehicle speed, subject to driver override, so as to maintain a safe distance from the nearest in-path vehicle ahead (the "lead" vehicle). If there are no lead vehicles within the stopping distance of the vehicle, the ACC functions as a conventional cruise control that maintains a fixed speed set by the driver. With lead vehicles present within the stopping distance, the system governs the acceleration and braking so as to control both the speed and the headway. Such radar has also been referred to by several other names such as intelligent cruise control (ICC), autonomous intelligent cruise control (AICC), and others.

Mission Requirements: First and foremost, the ACC radar must be capable of distinguishing between the closest vehicle ahead in the same lane and all other vehicles and roadside objects. As a result, a high accuracy in range and angular resolution is necessary. Second, it must acquire sufficient information to establish the minimum safe distance from the target vehicle, S_{min}. At the simplest level, S_{min} equals the stopping distance of the vehicle carrying the radar (the "host" vehicle), minus the stopping distance of the lead vehicle, along with an allowance for the distance traveled within the reaction time of the driver initiating the stopping action. If v_r and a_r are the velocity and deceleration of the host vehicle, v_t and a_t those of the targeted lead vehicle, and T_r the reaction time of the driver, then the minimum safe distance to the target can be estimated approximately as

$$S_{min} = \left(v_r^2/2a_r\right) - \left(v_t^2/2a_t\right) + v_r T_r$$

The distance calculated by this equation is subject to large uncertainty and additional safety margin, since the deceleration of each vehicle depends on the brake quality, road conditions, vehicle loading, and tire condition, while the reaction time depends on the driver's age, health, state of mind, training, and fatigue. However, the equation does show that to determine whether the following distance is safe requires not only the distance to the target but also the ground speed of the vehicle as well as the relative velocity with respect to the target. In particular, it is the radial component of the velocity that is pertinent, and the sign of the velocity is also important because it determines whether the vehicles are approaching or receding. This defines the minimum information that the ACC system must be designed to acquire.

16.6 Collision Anticipation Radar

Purpose: The purpose of collision anticipation radar is to sense an imminent collision. Several different variations of collision anticipation radars have been considered, differing in the use of the information gathered by the radar, and hence the definition of "imminent." For example, if the purpose of the radar is to initiate braking, the possibility of a collision must be sensed as early as possible to allow time for corrective action; if its purpose is to serve as a crash sensor for deploying an inflatable restraint system (commonly called airbag), only a collision that is certain is of interest. The different functions have been given various names, such as collision warning (CW), collision avoidance (CA), collision reduction (CR) radar, and others, but the nomenclature is not consistent, and is sometimes based on marketing rather than on technical differences. The following discussion illustrates some applications.

Collision Warning Application: Traffic accident analyses show that a significant fraction of traffic accidents (30% of head-on collisions, 50% of intersection accidents, and 60% of rear-end collisions) can be averted if the drivers are provided an extra 0.5 s to react. The purpose of the collision warning radar is to provide such an advanced warning to the driver. In order to perform that function, the radar must resolve, classify, and track multiple targets present in the environment; collect range, speed, and acceleration information about individual targets; use the past and current information to predict the vehicular paths over a short interval; estimate the likelihood of an accident; and present the situational awareness information to the driver through some human interface. Moreover, these functions must be performed in real time, within milliseconds, and repeated several times per second for updating.

Crash Sensing Application: Passive restraints (such as airbags and seat belt tensioners) used to protect vehicle occupants from severe injuries, typically employ several mechanical accelerometers to sense the velocity change in the passenger compartment. A radar used as an electronic acceleration sensor can have a number of advantages, such as sophisticated signal processing, programmability to customize it for each vehicle structure, self-diagnosis and fault indication, and data recording for accident reconstruction. Since the passive restraints require only about 30 ms to deploy, the ranges and time intervals of interest are shorter than in collision warning applications.

Radar Requirements: In each case, the collision anticipation radar must detect objects in the forward direction (in the path of the vehicle), and acquire range and velocity data on multiple targets. In this respect, the radar function is similar to that of ACC radar, and there are many similarities in the design considerations for the RF front end of the two types of radars. The system requirements and radar architecture for the two are therefore discussed together in the following.

16.7 RF Front-End for Forward-Looking Radars

ACC and collision anticipation radars have a number of similarities in the areas they monitor, information about the vehicular environment they must acquire, and the constraints under which they must operate. The major differences between them lie in the signal processing carried out on the radar return; the range of parameter values of interest; and the manner in which the collected information is utilized. From a user perspective, the primary difference between them stems from their roles: whereas the ACC radar is a "convenience" feature, a collision radar is thought as a "safety" device; consequently, the legal liability in case of malfunction is vastly different.

Radar Requirements: Some of the requirements to be satisfied by a forward-looking radar follow directly from its expected mission. The expected radar range is the distance to the lead vehicles of interest, and therefore lies in an interval of 3 m to perhaps as much as 200 m. The uncertainty in the measured value of range should not exceed 0.5 m. Since the lead vehicle is expected to be traveling in the same direction as the host vehicle, the maximum relative velocity of the vehicles can be expected to lie in the interval of +160 to –160 km/h. If the permissible uncertainty in the measurement of this relative speed is 1% at the maximum speed, a speed measurement error of up to 1.5 km/h is acceptable. If the information must be refreshed upon a change of 2 m in the range even at the highest speed, the radar needs to update the range and speed information once every 50 ms.

Environmental Complexity: The conditions under which the forward-looking radars operate is made complex by four features of the roadway environment. First, the radar must operate under harsh ambient conditions with respect to temperature, humidity, mechanical vibration and acceleration, and electromagnetic interference. Second, it must operate in inclement weather, caused by rain, sleet, snow, hail, fog, dust storm, and smoke. Third, the roadside scene is rapidly changing, and includes a large number of both potentially hazardous and harmless objects, some moving and others stationary, and the radar must identify and discriminate between them in order to maintain a low probability of false alarm. Fourth, due to road curvature and steering, and tangential components of vehicle velocities, the discrimination between in- and off-path objects becomes more involved, and requires computationally intensive prediction algorithms. Therefore, features like real-time signal processing, robust algorithms, and built-in fault detection, are essential.

Frequency Selection: Although some of the earlier radars were designed for operation around 16 and 35 GHz, virtually all current developments employ V-band frequencies for forward-looking radars. Within the V-band, several different frequencies have been used in the past decade, including 77 GHz for U.S. and European systems, and 60 GHz in some Japanese systems. Three factors dictate the choice of millimeter wave range for this application. First, the range resolution of the radar is governed primarily by the bandwidth of the transmitted signal, and a resolution of the order of 1 meter requires a minimum bandwidth of around 150 MHz. Such a large bandwidth is not available below the millimeter wave frequency range. Second, for a given performance level (such as antenna directivity), a higher frequency permits the use of smaller antennas and circuits, thereby reducing size, weight, and cost of the radar. Third, the higher atmospheric absorption of the millimeter wave signals is not a concern for short-range applications like automotive radars. However, the spray from other vehicles can impact the visibility of vehicles on the roadway at frequencies that lie in the water absorption band.

Signal Modulation: Although several different types of radar signal modulations have been evaluated over the years as possible candidates for forward-looking applications, most developers believe the frequency-modulated continuous-wave (FMCW) radar is the best overall choice. Some of the reasons for this choice of transmitted signal modulation include ease of modulation, simpler Doppler information extraction, higher accuracy for the short-range targets, and lower power rating for the active device generating the microwave signal for a given average transmitted power. In an FMCW radar, the frequency of the transmitted signal is changed with time in a prescribed manner (such as a linear ramp or a triangular variation), and the difference between the transmitted and returned signal frequencies is a measure of the range of the reflecting target.

Antenna Performance Requirements: Both the ACC and the collision anticipation radars require an antenna that meets several demanding specifications:

1. The size of the antenna should be small for cost and vehicle styling reasons.
2. The antenna beam should be narrow enough in azimuth that it can resolve objects in the driving lane from those in nearby lanes or adjacent to the roadway, as well as in elevation so as to distinguished on-road objects from overhead signs and bridges.
3. The sidelobes of the antenna should be sufficient low that small objects (like motorcycles) in the same lane are not masked by large objects (like trucks) in neighboring lanes.
4. The antenna should preferably be planar for ruggedness and ease of integration.

Choice of Antennas: Given the antenna beamwidth requirements and the limited physical dimensions permissible, some form of antenna scanning is needed for monitoring the relevant space, while at the same time maintaining the spatial discrimination. At least three choices are available for scanning. The first, and possibly the most versatile and powerful option is electronic scanning, using a phased array; with the presently available technologies, the cost of this option is prohibitive. Second, mechanically steered antennas have been developed that use reflectors. Third, synthetic aperture antenna techniques have been used that allow for sophisticated signal processing and information acquisition.

Signal Processing Needs: One of the most challenging aspects of automotive radar design, on which work presently continues, is the processing of the returned radar signals. Sophisticated signal processing algorithms have been employed to achieve many of the characteristics desired in radar performance, including the following:

1. Resolving small and large objects at different ranges and velocities.
2. Rejecting reflections from stationary objects.
3. Rejecting reflections from vehicles traveling in opposite directions.
4. Extracting target range despite road curvature.
5. Obtaining high performance despite low-cost components with low-performance.

This software must be capable of simultaneously separating and tracking a dozen or more different targets within the field of view.

The Radar Assembly: The need for large volume production and cost considerations dictate the materials, technologies, and processes usable in the radar. The complete radar can be conceptually subdivided into two parts: the RF part including the antenna, transmitter, and receiver, and the baseband part, consisting of signal processing, power supply, microprocessors, displays, cables, and the packaging and housing. The baseband part, as is typical of other automotive electronics, employs silicon chips, automated assembly methods such as flip-chip, and lightweight plastic housing. The cost of the front-end RF module at millimeter wave frequencies is usually minimized by use of a single hybrid assembly for the entire module, low-power Gunn devices as millimeter wave sources, silicon Schottky barrier diodes as mixers, and metalized injection-molded plastic waveguide cavities. More recently, as the cost of monolithic millimeter wave integrated circuit (MMIC) chips has decreased, a monolithic RF front end becomes cost effective, and can be integrated in a hybrid microstrip circuit. Chipsets for this purpose have been developed by a number of campanies.

16.8 Other Possible Types of Automotive Radars

Several other types of radar architectures and types have been proposed for automotive applications, and have reached different stages of development. Some of the promising candidates are briefly summarized here.

1. *Noise Radar.* Given the large number of automotive radars that may be simultaneously present in a given environment, the need to minimize the likelihood of false alarms due to interference between them is important. Several different types of noise (or noise-modulated) radars have been advanced for this purpose. One proposed scheme employs a transmitted signal modulated by random noise, and a correlation receiver that determines the cross-correlation between the returned signal modulation and the transmitted signal modulation, to separate the returned signal from the noise due to the system, ambient, and other radars. In still other schemes, the transmitted signal can be broadband noise, or a CW signal modulated by a binary random signal.
2. *Micropower Impulse Radar (MIR).* Another class of radar that has been proposed for several automotive uses is the so-called MIR. Its most distinguishing characteristic is that it transmits a very short pulse of electromagnetic energy, having a duration typically on the order of 0.1 ns, and a rise-time measured in picoseconds. As a result, its spectrum occupies a bandwidth of several GHz, creating a so-called "ultra-wideband" (UWB) system. Consequently, to avoid interference problems, the transmitted power is kept very low, on the order of a microwatt. As a further consequence of that choice, it is necessary to integrate over a large number of pulses to improve the signal-to-noise ratio of the receiver; however, the pulse repetition rate can be random, so that multiple radars will not interfere with each other due to their distinctive pulse patterns. In addition, the receiver is gated in time, so that it receives radar signal echoes only over a narrow time window. The received signal is thus limited to echoes from targets lying at a pre-selected distance from the transmitter, allowing the echoes from smaller nearby targets to be distinguished in the presence of those from large faraway objects. Another consequence of the low power transmission is long battery life, and hence the small volume, weight, and low cost. Some consequences of ultra-wideband operation are high range resolution, substantial scattering cross-section of targets at any observation angle, and signal penetration through dielectric materials.
3. *Interferometric Radar.* A radar system in which signals reflected from a target are simultaneously received by two physically separated antennas and are processed to determine the phase difference between them, can be used to estimate the target range over short distances. Such systems can employ a CW signal and a two-channel digital correlator, making them very simple, and can distinguish between approaching and receding targets. Since the range cannot be determined unambiguously from a known phase difference, the system is limited in its capability without signal processing to extract additional information from the returned signals.
4. *Cooperative Radar Systems.* Targets that modify the returned signal in known ways, for example, by modulation or frequency translation, to allow it to be easily detected and distinguished against other signals, are called cooperative targets. Radar systems with significantly higher performance capabilities can be developed if, as a result of common industry standards or regulatory directives, the vehicles incorporate appropriate means for enhancing returned radar signal, or carry a radar-interrogable distinguishing code, much like a license plate. In such a system, the ease of vehicle tracking can greatly improve the reliability and robustness of the radar system. The principal limitation of such sensors is their inability to handle targets with or without damaged reflectors, beacons, or other cooperative mechanisms.

16.9 Future Developments

Some of the major areas of emphasis in the future development of automotive radars are as follows:

1. *Radar Chipsets and MMICs*. With the frequency allocation and markets more definite, many manufacturers have developed chipsets for automotive radar that include GaAs MMIC chips for the front-end RF assembly. The availability of volume manufacturing capability, ease of incorporating additional functionality, and the resulting cost reductions make this a promising approach.
2. *Phased-Array Antennas*. With presently available technologies, phased-array antennas are not viable candidates for consideration in automotive radar applications due to their high cost. When low-cost phased-array antennas become commercially available, automotive radar can be expected to undergo a significant transformation and to attain a multifunction capability.
3. *Signal Processing Capability*. Digital processing of complex returned signals makes possible the characterization of more complex vehicular environments, in which large numbers of objects can be tracked. Improved algorithms, and the ability to carry out extensive real-time processing with inexpensive processor chips, will allow the radar to serve more sophisticated functions.
4. *Human Interface Enhancement*. The willingness of a driver to relinquish partial control of the vehicle to the radar depends only partly on attaining a high reliability, robustness, and low false-alarm probability. User acceptance will depend strongly on the quality of human interface through which the information gathered by an automotive radar is presented and utilized.

References

History

1. H. P. Groll and J. Detlefsen, History of automotive anticollision radars, *IEEE Aerospace and Electronic Systems Magazine*, 12, 8, 15–19, August 1997.
2. D. M. Grimes and T. O. Jones, Automobile radar: A brief review, *Proc. IEEE,* 62, 6, 804–822, June 1974.
3. M. S. Gupta, et al., Noise considerations in self-mixing IMPATT-diode oscillators for short-range Doppler radar applications, *IEEE Trans. Microwave Theory & Techniques*, MTT-22, 1, 37–43, January 1974.

Speed Measuring Radar

1. P. Heide, et al., A high performance multisensor system for precise vehicle ground speed measurement, *Microwave Journal*, 7, 22–34, July 1996.
2. P. Descamps, et al., Microwave Doppler sensors for terrestrial transportation applications, *IEEE Trans. Vehicular Technology*, 46, 1, 220–228, February 1997.
3. P. M. Schumacher, Signal processing enhances Doppler radar performance, *Microwaves & RF*, 30, 7, 79–86, June 1991.

Obstacle-Detection Radar

1. J. C. Reed, Side zone automotive radar, *IEEE Aerospace and Electronic Systems Magazine*, 13, 6, 3–7, June 1998.

Adaptive Cruise Control Radar

1. P. L. Lowbridge, Low cost millimeter-wave radar systems for intelligent vehicle cruise control applications, *Microwave Journal*, 38, 10, 20–33, October 1995.
2. L. H. Eriksson and S. Broden, High performance automotive radar, *Microwave Journal*, 39, 10, 24–38, October 1996.
3. M. E. Russel, et al., Millimeter wave radar sensor for automotive intelligent cruise control (ICC), *IEEE Trans. Microwave Theory & Techniques*, 45, 12, 2444–2453, December 1997.
4. A. G. Stove, Automobile radar, *Applied Microwaves*, 5, 2, 102–115, Spring 1993.

Collision Anticipation Radar

1. C. D. Wang and S. Halajian, Processing methods enhance collision-warning systems, *Microwaves & RF*, 36, 3, 72–82, March 1997.

RF Front-end for Forward-Looking Radars

1. W. H. Haydl, Single-chip coplanar 94-GHz FMCW radar sensors, *IEEE Microwave and Guided Wave Letters*, 9, 2, 73–75, February 1999.
2. D. D. Li, S. C. Luo, and R. M. Knox, Millimeter-wave FMCW radar transceiver/antenna for automotive applications, *Applied Microwaves and Wireless*, 11, 6, 58–68, June 1999.

Other Possible Types of Automotive Radars

1. Azevedo, S. and McEwan, T. E., Micropower impulse radar: A new pocket-sized radar that operates up to several years on AA batteries, *IEEE Potentials*, 16, 2, 15–20, April–May 1997.
2. G. Heftman, Macroapplications seen for micro radar, *Microwaves & RF*, 39, 5, 39–40, May 2000.
3. B. M. Horton, Noise modulated distance measuring systems, *Proc. IRE*, 47, 5, 821–828, May 1959.
4. I. P. Theron, et al., Ultrawide-band noise radar in the VHF/UHF band, *IEEE Transactions on Antennas and Propagation*, 47, 6, 1080–1084, 1999.
5. A. Benlarbi and Y. Leroy, A novel short-range anticollision radar, *Microwave and Optical Technology Letters*, 7, 11, 519–521, August 5, 1994.
6. F. Sterzer, Electronic license plate for motor vehicles, *RCA Review*, 35, 2, 167–175, June 1974.

Future Developments

1. Mueller, Jan-Erik GaAs HEMT MMIC chip set for automotive radar systems fabricated by optical stepper lithography, *IEEE Journal of Solid-State Circuits*, 32, 9, 1342–1349, September 1997.
2. Verweyen, L. Coplanar transceiver MMIC for 77 GHz automotive applications based on a nonlinear design approach, *1998 IEEE Radio Frequency Integrated Circuits Symposium Digest*, Baltimore, MD, June 1998, 33–36.
3. T. Shimura, et al., 76 GHz Flip-Chip MMICs for automotive radars, *1998 IEEE Radio Frequency Integrated Circuits (RFIC) Symposium Digest,* Baltimore, MD, June 1998, 25–28.

17

Ground Penetrating Radar

Carey M. Rappaport
Northeastern University

In many subsurface sensing and imaging applications, it is important to investigate noninvasively buried objects, underground structures, and the characteristics of soil and rocks below the ground surface. Along with seismic sounding, electromagnetic methods are the principle noninvasive means for gathering subsurface information about underground objects, including buried waste recovery, ground-water survey, land mine detection, and excavation planning. Commercial-off-the-shelf ground penetrating radar (GPR) subsurface detection systems are available for finding discrete buried objects and characterizing subsurface structures, and novel GPR designs have been developed for specific design challenges, such as mine and buried contaminant detection.

Almost all GPR systems are wide- or ultra-wideband radars, with specially designed, carefully matched wideband antennas. GPR systems can have an operating frequency from as low as 50 MHz to as high as 5 GHz. Lower frequencies are needed for deep wave penetration in wet, conductive soil, but high-resolution detection and imaging require higher probing frequencies.

Perhaps the greatest advantage GPR has over other subsurface sensing modalities is its ability to interrogate at distance. Ground contact tends to improve GPR sensing, but is not required. Standoff underground detection is critical for hazardous sensing applications, such as explosives detection, as well as for rapid sensor motion as when a pavement-monitoring GPR is mounted on a truck moving at highway speeds.

By carefully tailoring the transmitted signal and precisely measuring the received radar signal, several things can be inferred about the subsurface environment. In particular, the relative electrical characteristics of the scattering objects, their distances from the radar antenna, and their sizes can often be predicted with reasonable accuracy. The degree of wave scattering is directly related to the size and electromagnetic contrast of the scatterer. The more a region is electromagnetically different from its background and the larger it is, the more it scatters.

17.1 Special Aspects of Radar Wave Propagation in Soil

GPR differs from conventional radar in two major ways: the intended targets are completely or partially surround by media other than air, and near-field electromagnetic effects are paramount. These aspects make analysis and inversion of GPR signals challenging and quite different from conventional radar processing.

Soil is both lossy—meaning that waves attenuate, giving up power as they propagate through it, and dispersive—meaning different frequencies propagate with different velocity and decay rate. The wave equation for electric field E propagation in lossy dispersive media is

$$\nabla^2 E - \mu\sigma\frac{\partial E}{\partial t} - \mu\varepsilon\frac{\partial^2 E}{\partial t^2} = 0 \tag{17.1}$$

where μ, ε, and σ are the dielectric permittivity, magnetic permeability, and electrical conductivity, respectively. Typical values for these parameters are $\mu \approx \mu_0$, $\varepsilon_0 < \varepsilon < 25\varepsilon_0$, $0.0001 < \sigma < 1$. Converting to the frequency domain, this equation becomes

$$\nabla^2 E + k^2 E = 0 \tag{17.2}$$

The dispersion relation relating wave number to angular frequency $\omega = 2\pi f$ is given by

$$k = \omega\sqrt{[\mu_0\varepsilon_0\varepsilon'(1 - j\tan\delta)]} = \beta - j\alpha \tag{17.3}$$

where ε' is the relative permeability, and the loss tangent is defined by

$$\tan\delta = \frac{\sigma}{\omega\varepsilon'\varepsilon_0} \tag{17.4}$$

The plane wave solution to Equation 17.2 in the frequency domain is

$$E(x, t) = \mathrm{e}^{-j[\beta(f) - j\alpha(f)]x}\,\mathrm{e}^{j2\pi ft} \tag{17.5}$$

for a wave propagating in the $+x$ direction. Although there are commonly used approximations for weakly and strongly conducting media, soil tends to have intermediate electrical conductivity, necessitating the full formulation for the propagation number and decay rate in terms of frequency:

$$\beta = \frac{\omega}{c}\left\{\frac{\varepsilon'}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon'\varepsilon_0}\right)^2} + 1\right]\right\}^{1/2} \tag{17.6}$$

$$\alpha = \frac{\omega}{c}\left\{\frac{\varepsilon'}{2}\left[\sqrt{1 + \left(\frac{\sigma}{\omega\varepsilon'\varepsilon_0}\right)^2} - 1\right]\right\}^{1/2} \tag{17.7}$$

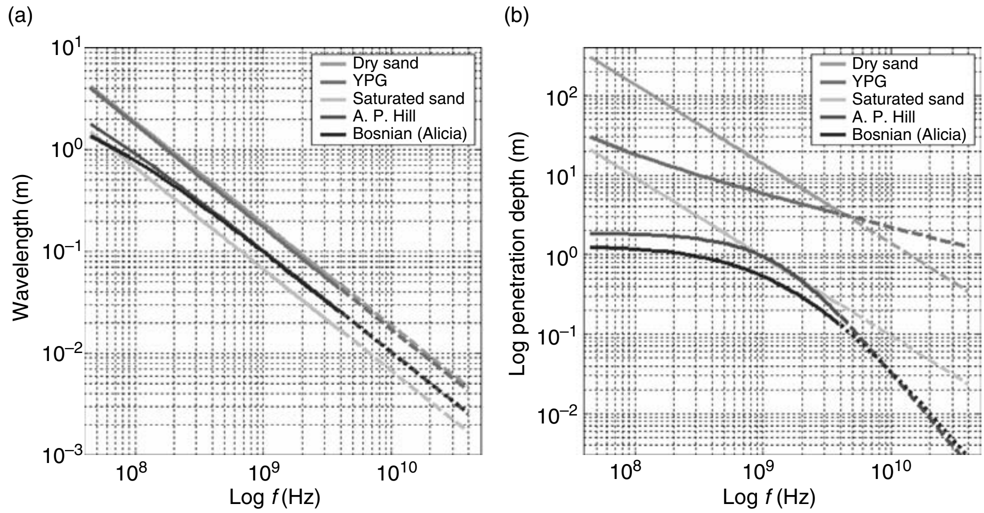


FIGURE 17.1 (a) Wavelength and (b) penetration depth for dry and saturated sand, Yuma Proving Ground soil, 19% moist Ft. A. P. Hill clay loam, and 26% moist Bosnian loam.

The propagation velocity, wavelength, and −10 dB penetration depth (distance required for power to drop by a factor of 10) are given in terms of the propagation number by

$$\begin{aligned} v &= \omega/\beta \\ \lambda &= 2\pi/\beta \\ d_{10} &= \frac{\log_e 10}{2\alpha} \end{aligned} \tag{17.8}$$

Values of wavelength and −10 dB penetration depth for typical soils are shown in Figure 17.1.

The great variation in these parameters indicates that soil type and moisture determine the range of applicability of radar in penetrating ground. Since imaging resolution is roughly limited to one-half of a wavelength, Figure 17.1a indicates the minimum necessary frequency to distinguish a particular target for a given soil type. The typical GPR dynamic range and clutter rejection limits two-way loss to a maximum of about 50 dB, which sets the maximum useful frequency for a given target depth.

17.2 Soil Reflection Coefficient and Wave Polarization

The effectiveness of GPR in observing below-the-ground surface is dependent on how much of the electromagnetic wave penetrates into the ground, and how much is reflected back into the air. The reflection coefficient is mainly a function of the wave incident angle and the frequency-dependent dielectric characteristics of the ground. For transverse electric (TE) waves, which have the electric field pointing parallel to the ground, the reflection coefficient increases from its lowest value at normal incidence to 100% reflection for fully grazing incidence. A special feature of the perpendicular transverse magnetic (TM) polarization is that the reflection decreases with incidence angle until it reaches a minimum at the so-called Brewster angle. If the half space is lossless, 100% transmission occurs at the Brewster angle. For conductive soil, the reflection coefficient is smallest, but nonzero. Figure 17.2 shows reflection coefficients as a function of incidence angle for typical dielectric soil values. Forward-looking GPR, used in sensing systems with large standoff, often use TM polarized waves to take advantage of the Brewster angle maximum transmission phenomenon to direct most of the wave power into the ground.

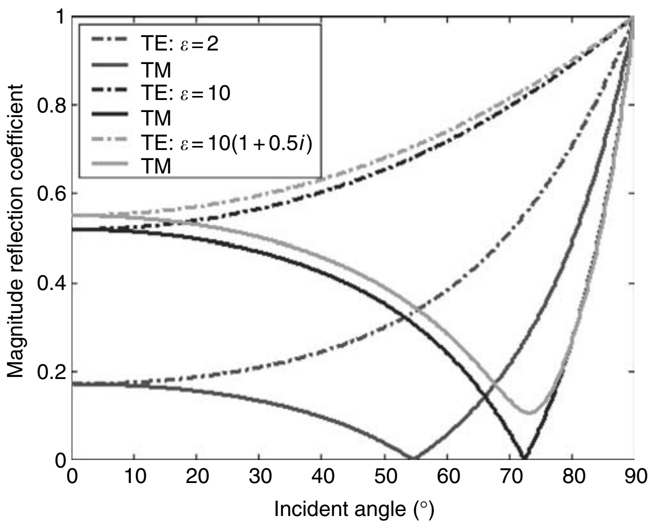


FIGURE 17.2 Reflection coefficients for TE and TM plane waves from lossless and lossy half spaces.

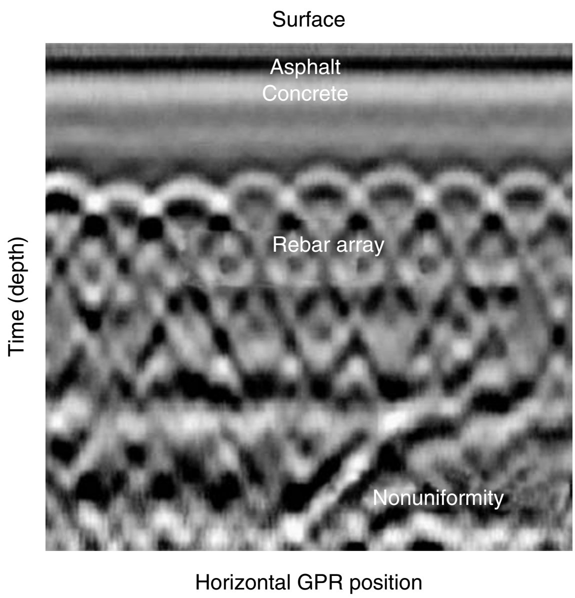


FIGURE 17.3 B-scan of pavement with parallel layers of asphalt, concrete, and rebar array.

17.3 Impulse and Stepped Frequency GPR

GPR systems are divided into two main types, impulse and stepped frequency. Impulse GPR systems operate by transmitting a short microwave pulse into the ground and observing the returning signal scattered by target objects. Wideband radar components are used to generate, transmit, and receive the pulses with the shortest possible rise and fall times (of the order of 30–100 ps), with minimal ringing. The amplitude, time delay, and shape of the returned time pulse are measured and processed to infer subsurface structure. Common representations of radar response are time traces for a given GPR position, and stacked time traces for various GPR positions, referred to as B-scans. Figure 17.3 shows a B-scan of pavement with an array of reinforcing steel bars. The time responses of received signals are displayed as

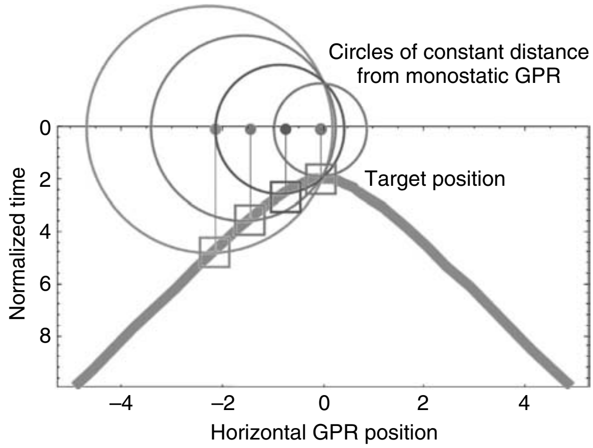


FIGURE 17.4 Generating the B-scan hyperbolic time contour from various GPR positions on the ground surface (dots), recording the time to/from the target as the radius of the circle intersecting the target at $(0, -d)$, with distances and velocity normalized by target depth d.

shades of gray, for each horizontal GPR position, with the time axis increasing downward. Clearly visible in this image are the reflection from the top of the concrete pavement, and eight overlapping downward opening hyperbolas.

The characteristic hyperbolic shape, one for each metal bar, is the primary feature of impulse GPR response from individual buried objects. Figure 17.4 schematically demonstrates how the hyperbola is formed. Given a scattering target depth d in a uniform half space with wave propagation velocity $v = c/\sqrt{\varepsilon'}$ in soil with relative permeability ε', a horizontal GPR position x relative to the point directly above the target, the two-way travel time from the GPR to the target is $T = 2\sqrt{x^2 + d^2}/v$, or in standard form

$$\left(\frac{T}{2}\right)^2 - \left(\frac{x}{v}\right)^2 = \left(\frac{d}{v}\right)^2$$

which is the equation of a hyperbola, with slope v, and vertex d/v down. This is shown in Figure 17.3, with velocity v normalized to target depth d.

Stepped frequency systems transmit continuous wave signals at individual frequencies varying through a particular band and receive the amplitude and phase of the scattered signal at each separate frequency. The frequency domain information is then assembled into a frequency response for the sample region. Advantages over impulse systems include the ability to excite or suppress—as well as quickly process—specific frequency ranges, minimal difficulty with signal generation, the full characterization of the transmitting and receiving properties of the antennas, and the availability of high resolution data for specific frequencies. Disadvantages include higher cost, longer data acquisition and operation time, and greater processing required to produce easily interpretable information. Tomographic imaging systems require phase as well as amplitude, and are best suited for stepped frequency GPR.

Figure 17.5 shows the real part of the spatial field distribution of the total electric field due to a 1 GHz point source above a soil half space with a planar boundary. The transmitted waves have shorter wavelength, corresponding to the reduced velocity in the soil, and they clearly attenuate with distance into the soil. The reflected waves interfere with the incident waves in the air above the ground interface at height 0. The waves progress in direction perpendicular to the wavefronts along rays, several of which are shown in Figure 17.5. The rays indicate the expected refraction at the planar ground interface following Snell's Law $\sqrt{\varepsilon_{\text{soil}}}\sin\theta_t = \sqrt{\varepsilon_0}\sin\theta_i$. Note that the wavelength in soil λ_g is reduced by the ratio $\sqrt{\varepsilon_0/\varepsilon_{\text{soil}}}$ from the wavelength in air λ_0.

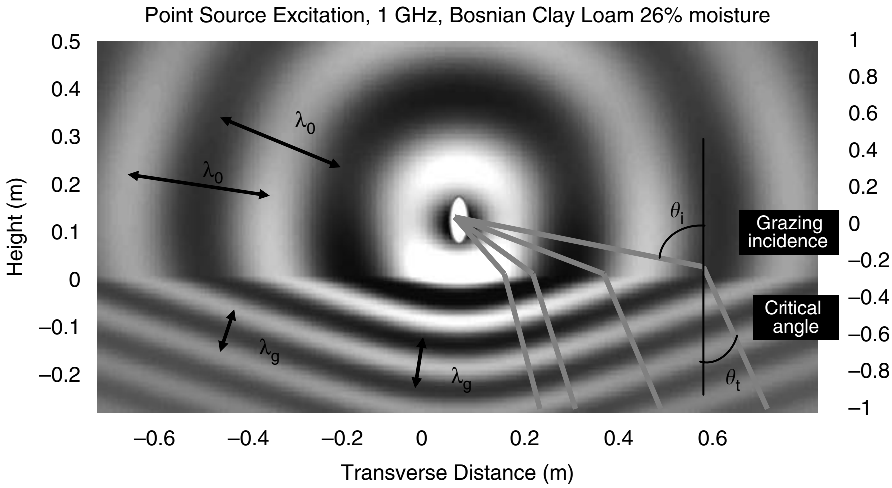


FIGURE 17.5 Computed real part of electric field due to a single frequency point source directed perpendicular to the image plane, above a lossy soil half space.

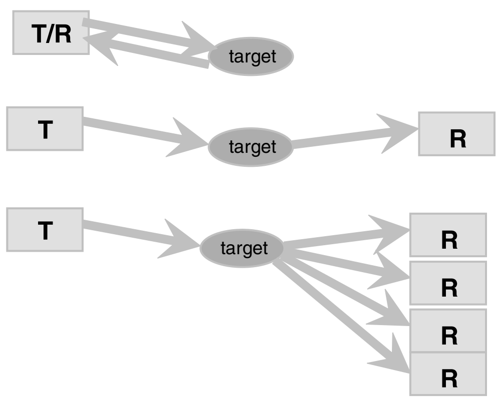


FIGURE 17.6 Monostatic, bistatic, and multistatic radar configurations.

17.4 Monostatic, Bistatic, Multistatic, and Synthetic Aperture Radar GPR

A GPR can have various transmitter and receiver configurations. If the same antenna is used for both, it is referred to as monostatic. Separated single transmitter/receiver pairs are the bistatic arrangement; and multiple receivers or transmitters constitute the multistatic configuration. Multiple views of a given scene by a single antenna GPR is often describes as multimonostatic radar. Figure 17.6 schematically shows the various combinations.

Monostatic GPR is the least expensive and the simplest to use, as it is easy to position (and record its position), there is no relative motion between the transmitting and receiving antenna, and the impedance matching for a single antenna avoids measurement uncertainties inherent in viewing different regions of ground with multiple antennas. Monostatic systems do require high performance transmit/receive separation circuitry to avoid saturating/overwhelming the receiver with the strong transmitted signal. The chief disadvantage of the monostatic configuration is the limited available aperture and the strong correlated effects of surface roughness clutter. Bistatic and monostatic systems view the target from different positions, often providing the processing algorithms with information to reduce this surface clutter. In addition, separated transmitters and receivers can give better target position information. For bistatic and multistatic GPR, the large direct signal must be calibrated and removed to record the much smaller ground scattered signal. In addition, very accurate timing and positioning between transmitter and receiver is necessary. Multistatic systems make use of the largest possible aperture, offering the best potential for ground surface clutter rejection. Multistatic GPR is necessary for tomographic subsurface reconstruction, in which the simultaneous scattered views of the target are combined into a cross-sectional image. Tomography will be discussed in a subsequent section.

Multimonostatic GPR involves multiple gathers of monostatic information, with a repeated sequence of acquisition and reposition. The motion of the radar antenna in time provides a large aperture, with data from each return combined with others to form an image (i.e., a B-scan for impulse GPR). One particular advantage of multimonostatic radar is that small repositioning can provides dense sampling, which would be unavailable with a single large aperture. With processing to combine the stacked repositioned traces into a single response due to an artificially created large aperture, multimonostatic radar becomes synthetic aperture radar (SAR). Traditional SAR greatly improves the gain and cross-range resolution for airborne radars, but it is not as effective for GPR in which the ground surface wave refraction complicates the combining of multiple monostatic measurements.

17.5 Detecting Penetrable Dielectric Objects with GPR

While buried metallic targets are the most readily detected underground objects, nonmetallic objects with sufficiently large contrast scatter microwaves as well, and can often be detected and imaged. GPR has proved valuable for archeological and geophysical surveying, underground contaminant, and structure mapping, and for unexploded ordnance and landmine detection.

For example, to find buried mines, the GPR system focuses waves on points below the ground surface, and then uses the received scattered waves to quantify the electromagnetic anomalies in an inhomogeneous soil background environment. Unlike sensing defects in regular underground structures, such as road pavement or bridge decks, detecting isolated targets in nonuniform soil often requires extensive processing to remove clutter, to compensate for refraction through rough interfaces, and to reconstruct the most probable target size and configuration. Small plastic antipersonnel (AP) mines buried only a few inches below the ground surface are challenging targets, but since their dielectric composition and expected shape are known, many signal processing methods can be used to discriminate mine targets from natural and manmade clutter.

It is essential that the soil dielectric characteristics are considered when analyzing the GPR response from dielectric targets. Figure 17.7 shows the modeled scattered field from the same spherical TNT target in three different background half spaces: air, dry sand, and moist clay loam soil. In each case, the GPR antenna is assumed to be located in air above the half space. Figure 17.7a shows the field distribution if the sphere were simply floating in air, corresponding to conventional radar sensing. Note that this distribution is completely different, both viewed in cross section and at the nominal surface, from the sand and soil background cases, Figure 17.7b and c. In addition, the dielectric constant differences between sand and soil generate remarkable differences in these two cases.

The most challenging problem with GPR, however, is detecting an object right at, or just below the ground surface. This is because the reflection from the air/ground interface is quite strong to begin with, and the smaller scattered signal from the near-surface target would occur at almost the same time as the ground reflection signal. In effect, the target signal would be overwhelmed by the surface reflection from the ground.

17.6 Modeling GPR Wave Scattering and Inversion Methods

For sophisticated quantitative analysis of the subsurface, it is essential to model the electromagnetic interaction of microwaves with underground objects and structures. Computational models provide a

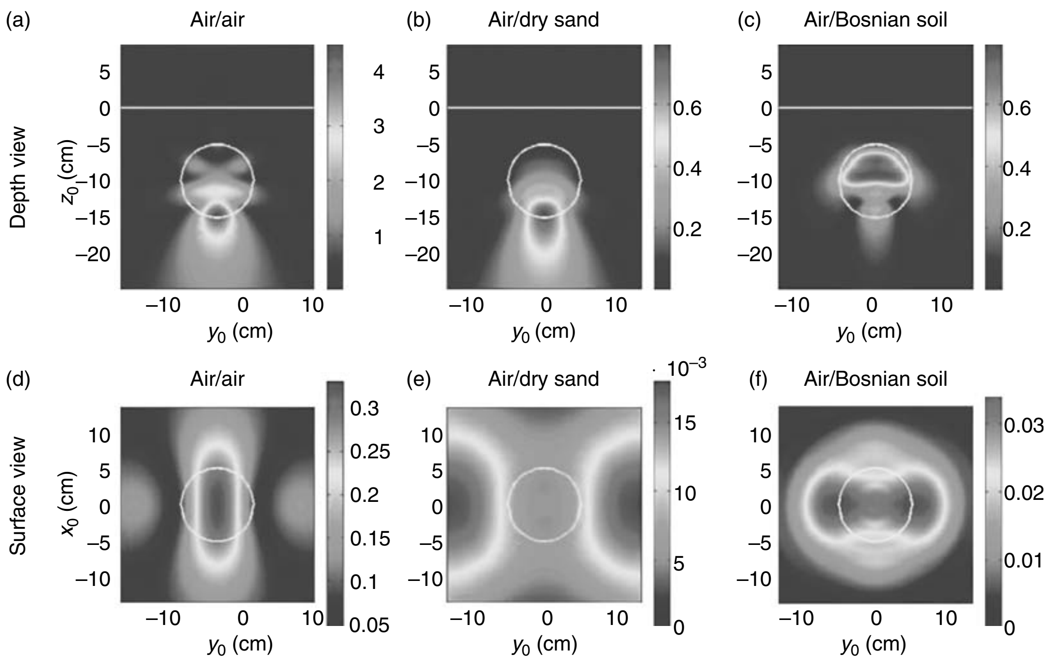


FIGURE 17.7 Electric field magnitude scattered by a spherical object with the dielectric characteristics of TNT: $\varepsilon = 2.9\varepsilon_0$, $\sigma = 0.001$ S/m, for a 3 GHz normally incident plane wave from air onto three background half-spaces: air (a) depth view (d) surface view; dry soil (b) depth view (e) surface view; and Bosnian clay loam (c) depth view (f) surface view.

basis for identifying target signal features, guide the placement of antennas, and serve as the forward models for inversion analysis and reconstruction.

To model the subsurface scattering process correctly it is also necessary to accurately determine the electromagnetic fields refracting through the ground interface. In nature, the soil surface is usually rough, with random variation of surface height and peak separation. Since the degree of roughness is often a significant fraction of the length scale of the sensing geometry—the target size, its buried depth, and the field wavelength—it is essential to incorporate the estimates of parameters of this medium interface in field modeling. In addition, since the scattering of near-surface objects occurs in the near-field of GPR antenna systems, high-resolution methods are necessary.

Only a limited number of models are available for computational modeling of the field generated by a realistic GPR both above and below a rough ground surface. These options include the finite difference time domain (FDTD) and frequency domain (FDFD) methods, the method of moments (MoM), the finite element (FE) method, the modal expansion methods, such as T-Matrix and the semi-analytic mode matching method (SAMM), and Gaussian beam (GB) methods. Each of the methods solves Maxwell's equations applied to the problem space by dividing the space or the form of the solution into small, manageable pieces for which approximate solutions are possible. With spatial discretization, either the scatterers or the entire problem domain is represented by a large ensemble of subsectional regions. For the modal and beam methods, the total field is decomposed into a series of simpler piecewise field solutions, which are then combined such that they obey boundary conditions.

While some buried target detection applications rely solely on a change in the relatively constant GPR return as it is moved over the ground, it is often possible to use GPR signals to characterize or even image the underground objects. Using the measured received signals to reconstruct the subsurface nonuniformities is referred to as "inversion." For simple shapes or structures, multiple surface measurements can provide sufficient information to predict their positions, sizes, and aspect ratios.

Inversion is complicated by several factors. The background environment can be extremely varied and heterogeneous, so it is hard to separate the target characteristics from the background. The targets may come in many different types and sizes, and the distance to target might be large compared to target size and depth. The contrast of nonmetallic objects might be weak, so that its scattering is small compared to the scattering from surface roughness or background inhomogeneities.

The clutter that obscures and confuses the target signal is due mostly from surface reflection. The correlated errors from random ground surface variations makes the effects on the target signal hard to predict on an individual view basis. The global effects of clutter are correlated with statistics of random surface variations.

For impulse GPR, two main inversion methods are used: migration and time reversal. Migration involves converting a B-scan response taken on the ground surface as a function of time and horizontal position x into a spatial image as a function of horizontal position and depth z, taken at an initial time $t = 0$. This is accomplished by transforming into the frequency ω and spatial-frequency k_x, k_z space, and using the dispersion relation for the propagation medium. First, the time domain B-scan signal gathered at the ground surface $f(t, x, z = 0)$ is Fourier transformed, both in time and space, into a frequency domain/spatial-frequency domain signal $F(\omega, k_x, z = 0)$. Next, the phase is retarded by multiplying by $\exp(-jk_z z)$, and the dispersion relation $\omega = v\sqrt{k_x^2 + k_z^2}$ is used to substitute for the frequency variable to convert the field representation into a function of just spatial-frequency variables $F(k_x, k_z)$. Finally, this last function is doubly inverse Fourier transformed back to the physical space and time is set to zero, $t = 0$, resulting in $f'(x, z, t = 0)$. This gives the image of the subsurface structure along the plane defined by the transverse motion of the GPR and depth into the ground. Certain modifications are necessary when dealing with layered geometries, including the case when the GPR antenna is in air, above the ground surface.

Time reversal is a useful time domain inversion technique for imaging a small number of isolated targets. It is particularly well suited for high-contrast scatterers. The general idea is to compute the wave propagation backward. Simplistically, this can be done by observing expanding wavefronts, calculating their normal rays, and tracing these back—with appropriate refraction at known layers to a single source of the scattering. More effective is to record the full time signals on a densely sampled grid above the target region, and use FDTD with a negative time step to backpropagate the wave, stopping at the time instant when the greatest field refocusing occurs. The position of highest intensity is the best prediction of scattering source location. Time reversal concepts can also be extended into frequency domain systems when the background medium can be approximated as lossless.

Inversion techniques in the frequency domain are more common for microwave medical and ultrasound imaging modalities, but have also been applied to stepped frequency GPR. The most common form of single frequency inversion is tomographic. Tomography generates an image of the entire illuminated region, making it useful for characterizing distributed variations of material type. Since the inversion is conducted at a single frequency, many observations of wave amplitude and phase are required. For noninvasive above-surface GPR, only backscatter measurements can be made, which limits the resolving capability of the tomographic inversion. Cross-borehole GPR measures transmission as well as reflection signals, and thus is better suited to frequency domain inversion.

Tomographic backpropagation inversion makes use of the Born or Rytov approximations to simplify the integral equation for describing the scattered field in terms of unknown scattering variations in the background medium. Since they neglect multiple interactions, these approximations are best when the contrasts and sizes of the variations are small. Conceptually, the integral equation represents the sum of each infinitesimal scattering source contribution, each excited by a known incident field. Usually, these contributions are determined using the uniform space Green's function, which is just the field due to a Hertzian dipole point source in an infinite medium. Half space Green's functions are also used for GPR inversion, but are much more complicated. To compute the image of the contrast variations, the image space is discretized (represented by the vector $\bar{x}$ of unknown pixel dielectric constant values), along with the known measured observation space $\bar{b}$, and Green's function that associates points in the image space to observation space $\overline{\overline{A}}$. The unknown pixel values of the image are then just the solution to the matrix equation $\overline{\overline{A}} \cdot \bar{x} = \bar{b}$.

Bibliography

1. Birsen, C., "Dielectric Properties of Mixtures of Clay-Water-Organic Compounds," PhD Thesis, Colorado School of Mines, 1999.
2. Black, C.A., *Methods of Soil Analysis: Part I Physical and Mineralogical Properties*, Madison, WI: American Society of Agronomy, 1965.
3. Blitz, J., *Electrical and Magnetic Methods of Nondestructive Testing*, Bristol, Philadelphia: A. Hilger, 1991.
4. Bray, D.E., and McBride, D., *Nondestructive Testing Techniques*, New York: Wiley, 1992.
5. Carin, L., Kapoor, R., and Baum, C., "Polarimetric SAR imaging of buried landmines," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 36, No. 6, 1998, pp. 1985–1988.
6. Cole, K.S., and Cole, R.H., "Dispersion and absorption in dielectrics," *Journal of Chemical Physics*, Vol. 9, 1941, pp. 341–351.
7. Dogaru, T., Collins, L., and Lawrence Carin, L., "Optimal time-domain detection of a deterministic target buried under a randomly rough interface," *IEEE Transactions on Antenna and Propagation*, Vol. 49, No. 3, 2001, pp. 313–326.
8. Dorn, O., Miller, E., and Rappaport, C., "A shape reconstruction method for electromagnetic tomography using adjoint fields and level sets," *Inverse Problems*, Vol. 16, 2000, pp. 1119–1156.
9. Dumanian, A., Rappaport, C., and Morgenthaler, A., "Time domain processing of frequency domain GPR signatures for buried land mine detection," *Proceedings of the SPIE Aerosense Conference*, Orlando, FL, April 2002, pp. 339–348.
10. Dumanian, A., and Rappaport, C., "Enhanced detection and classification of buried mines with an UWB multistatic GPR," *Proceedings of the IEEE International Symposium on Antennas and Propagation*, Washington, D.C., July 2005, four pages on CD.
11. El-Shenawee, M., Rappaport, C., Miller, E., and Silevitch, M., "3-D Subsurface analysis of electromagnetic scattering from penetrable/PEC objects buried under rough surfaces: Application of the steepest descent fast multipole multilevel (SDFMM)," *IEEE Transactions on Geoscience and Remote Sensing*, June 2001, pp. 1174–1182.
12. El-Shenawee, M., and Rappaport, C., "Modeling clutter from Bosnian and Puerto Rican rough ground surfaces for GPR subsurface sensing applications using the SDFMM," *International Journal of Subsurface Sensing Technologies and Applications*, July 2001, pp. 249–264.
13. El-Shenawee, M., Rappaport C., Miller, E., and Silevitch M., "Three-dimensional subsurface analysis of electromagnetic scattering from penetrable/PEC objects buried under rough surfaces: use of the steepest descent fast multipole method (SDFMM)," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 39, No. 6, 2001, pp. 1174–1182.
14. El-Shenawee, M., Rappaport, C., and Silevitch, M., "Monte Carlo simulations of electromagnetic wave scattering from random rough ground surface with 3-D penetrable buried object: Mine detection application using the SDFMM," *Journal of the Optical Society of America*, Vol. 18, No. 12, 2001, pp. 3077–3084.
15. El-Shenawee, M., Rappaport, C., Zhan, H., and Miler, E., "Mine detection under rough ground surfaces using 2-D FDTD modeling and hypothesis testing," *Proceedings of the IEEE Antenna and Propagation Society International Symposium*, Boston, MA, Vol. 3, 2001, p. 756.
16. Farid, M., Alshawabkeh, A., Zhan, H., and Rappaport, C., "Challenges and validation of cross tomography experimentation for inverse scattering problems in soil," *Proceedings of Geo-Frontiers 2005: GSP 138, Site Characterization and Modeling*, Austin, TX, 25–27 January 2005, Ten pages on CD.
17. Galdi, V., Kosmas, P., Rappaport, C., Felsen, L., and Castanon, D., "Short pulse three-dimensional scattering from moderately rough surfaces: A comparison between narrow-waisted Gaussian beams and FDTD," *IEEE Transactions on Antenna and Propagation*, Vol. 54, No. 1, 2006, pp. 157–167.
18. Gardner, C., Robinson, D., Blyth, K., and Cooper, J.D. "Soil water content," In K.A. Smith and C.E. Mullins (ed.), *Soil and Environmental Analysis: Physical Methods*, 2nd ed., New York: Marcel Dekker, 2001, pp. 1–64.

19. Gazdag, J., and Sguazzero, P., "Migration of seismic data," *Proceedings of the IEEE*, Vol. 77, No. 10, 1984, pp. 1302–1315.
20. Geng, N., and Carin, L., "Wide-band electromagnetic scattering from a dielectric BOR buried in a layered lossy dispersive medium," *IEEE Transactions on Antenna and Propagation*, Vol. 47, No. 4, 1999, pp. 610–619.
21. Gibson, P.J., and George, D.M., *Environmental Applications of Geophysical Surveying Techniques*, Hauppauge, NY: Nova Science Publishers, 2004.
22. Hipp, J.E., "Soil electromagnetic parameters as functions of frequency, soil density and soil moisture," *Proceedings of the IEEE*, Vol. 62, No. 1, 1974, pp. 98–103.
23. Hoekstra, P., and Delaney, A., "Dielectric properties of soils at UHF and microwave frequencies," *Journal of Geophysical Research*, Vol. 74, 1969, pp. 5867–5881.
24. Holtz, R.D., and Kovacs, W.D., *An Introduction to Geotechnical Engineering*, Englewood Cliffs, NJ: Prentice-Hall, 1981.
25. Huisman, J.A., Hubbard, S.S., Redman, J.D., and Annan, A.P., "Measuring soil water content with ground penetrating radar: a review," Department of Landscape Ecology and Resources Management, Justus-Liebig-University, Giessen, Germany.
26. Ida, N., *Microwave NDT*, Dordrecht; Boston: Kluwer Academic, 1992.
27. Johansen, P., Rappaport, C., Devaney, A., and Miller, E., "Electromagnetic inversion for multi-static ground penetrating radar," *Progress in Electromagnetic Research Symposium*, July 1998, p. 862.
28. Koh, G., "Effect of soil moisture on radar detection of simulant mines," fact sheet, *Hanover, N.H.: U.S. Army Corps of Engineers, Cold Regions Research, and Engineering Laboratory*, 1998.
29. Kosmas, P., Wang, Y., and Rappaport, C., "A three-dimensional FDTD model for GPR detection of objects buried in realistic dispersive soil," *Proceedings of the SPIE Aerosense Conference*, Orlando, FL, April 2002, pp. 330–338.
30. Kraus, J., *Antennas*, New York: McGraw-Hill, 1988.
31. Kunz, K.S., and Luebbers, R.J., *The Finite Difference Time Domain Method for Electromagnetics*, London: CRC Press, 1993.
32. Lambot, S., Slob, E.C., den Bosch, I.V., Vanclooster, M., Stockbroeckx, B., and Druyts, P., "Frequency dependence of the soil electromagnetic properties derived from ground-penetrating radar signal inversion," *Subsurface Sensing Technologies and Applications*, Vol. 6, No. 1, 2005, pp. 73–87.
33. MacDonald, J., Lockwood, J., McFee, J., Altshuler, T., Broach, T., Carin, L., Harmon, R., Rappaport, C., Scott, W., and Weaver, R., *Alternatives for Landmine Detection*, Santa Monica: RAND, 2003.
34. Martinez, A., and Byrnes, A.P., *Modeling Dielectric-Constant Values of Geologic Materials: An Aid to Ground-penetrating Radar Data Collection and Interpretation*, Kansas Geological Survey, The University of Kansas, Lawrence, 2001.
35. Morgenthaler, A., and Rappaport, C., "Scattering from lossy dielectric objects buried beneath randomly rough ground: validating the semi-analytic mode matching algorithm with two-dimensional FDFD," *IEEE Transactions on Geoscience and Remote Sensing*, November 2001, pp. 2421–2428.
36. Noborio, K., "Measurement of soil water content and electrical conductivity by time domain reflectometry: a review," *Computers in Electronics and Agriculture*, Vol. 31, 2001, pp. 213–237.
37. Olhoeft, G. *Ground Penetrating Radar: Introduction and History*, Available at http://www.g-p-r.com/introduc.htm
38. Olhoeft, G.R., and Capron, D.E., "Laboratory measurements of the radio frequency electrical and magnetic properties of soils near Yuma, Arizona," US Department of Interior USGS Open File Report 93-701, 1993.
39. Parrillo, R., Roberts, R., and Delea, D., "Comparison of 2 GHz horn antenna and 1.5 GHz ground-coupled antenna for bridge deck condition assessment using GPR," *Structural Faults & Repair*, Edinburgh ECS Publication, June 2006.
40. Rappaport, C., Azevedo, S., Rosenbury, T., and Gough, J., "Handheld forward-looking focused array mine detection with plane wave excitation," *Proceedings of the SPIE Aerosense Conference*, Orlando, FL, April 2000, pp. 1118–1126.
41. Rappaport, C., El-Shenawee, M., and Zhan, H., "Suppressing GPR clutter from randomly rough ground surfaces to enhance mine detection," *International Journal of Subsurface Sensing Technologies and Applications*, Vol. 4, No. 4, 2003, pp. 309–324.

42. Rappaport, C., "Exact and approximate underground wavefront shape determination from an above ground GPR source," *IEEE International Symposium on Antennas and Propagation*, Washington, D.C., July 2005, Four pages on CD.
43. Rappaport, C., Miller, E., and Kilmer, M., "A new shape-based method for object localization and characterization from scattered field data," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 38, No. 4, 2000, pp. 1682–1696.
44. Rappaport, C., "Modeling GPR surface roughness clutter effects for mine detection," in *Ground Penetrating Radar*, D. Daniels (ed.), UK: IEE Press, 2004, pp. 58–72. *004 URSI Conference*, Pisa, Italy, 16–18 May, 2004, pp. 1134–1136.
45. Rappaport, C., "Surface Roughness Clutter Effects in GPR Modeling," in *Wave Propagation, Scattering, and Emission in Complex Media*, Y.Q. Jin (ed.), Beijing: Science Press, 2004, pp. 119–127.
46. Rappaport, C., Winton, S., Jin, D., and Siegle, L., "Modeling the effects of non-uniform soil moisture on detection efficacy of mine-like objects with GPR," *SPIE Aerosense Conference*, Orlando, FL, April 1999, pp. 979–990.
47. Rappaport, C., Wu, S., Kilmer, M., and Miller, E., "Distinguishing shape details of buried non-metallic mine-like objects with GPR," *Proceedings of the SPIE Aerosense Conference*, Orlando, FL, April 1999, pp. 1419–1428.
48. Santamarina, J.C., Klein, K.A., and Fam, M.A., *Soils and Waves: Particulate materials Behavior, Characterization and Process Monitoring*, New York: Wiley, 2001.
49. Scott, M., Rezaizadeh, A., Delahaza, A., Santos, C, Moore, M., Graybeal, B., and Washer, G., "A comparison of nondestructive evaluation methods for bridge deck assessment," *NDT&E International*, Vol. 36, 2003, pp. 245–255.
50. Taflove, A., *Computational Electrodynamics—The Finite Difference Time-Domain Method*, London: Artech House, 1993.
51. Topp, G.C., Davis, J.L., and Annan, A.P., "Electromagnetic determination of soil water content: Measurement in coaxial transmission lines," *Water Resources*, Vol. 16, 1980, pp. 574–582.
52. Trabelsi, S., Kraszewski, A.W., and Nelson, S.O., "Microwave dielectric sensing of bulk density of granular materials," US Department of Agriculture, Agricultural Research Service, Quality Assessment Research Unit, Athens, GA, USA, 2001.
53. van Dam, R.L., Borchers, B., and Hendrickx, J.M.H., "Methods for prediction of soil dielectric properties: A review," in *Detection and Remediation Technologies for Mines and Minelike Targets X*, Vol. 5794, Orlando, FL: SPIE, 2005, pp. 188–197.
54. von Hippel, A.R., *Dielectrics Materials and Applications*, Cambridge, MA: MIT Press, 1954.
55. Xu, X., Miller, E., Rappaport, C., and Sower, G., "Statistical method to detect subsurface objects using array ground penetrating radar data," *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 40, No. 4, 2002, pp. 963–976.
56. Yang, B., and Rappaport, C., "Response of realistic soil for GPR applications with two dimensional FDTD," *IEEE Transactions on Geoscience Remote Sensing*, 2001, pp. 1198–1205.
57. Yee, K.S., "Numerical solution of initial boundary value problems involving Maxwell's equation in isotropic media," *IEEE Transactions on Antennas Propagation*, Vol. 14, No. 2, 1966, pp. 302–307.
58. Zhan, H., Rappaport, C., Farid, M., Alshawabkeh, A., and Raemer, H., "Lossy half-space Born approximation modeling of electromagnetic wave source and scattering in soil by cross well radar," *Proceedings of Geo-Frontiers 2005: GSP 138, Site Characterization and Model*, Austin, TX, 25–27 January 2005, Ten pages on CD.

18

New Frontiers for RF/Microwaves in Therapeutic Medicine

Arye Rosen
Drexel University

Harel D. Rosen
Onsite Neonatal Partners, Inc.

Stuart D. Edwards
Silhouette Medical

The use of RF/microwaves in therapeutic medicine has increased dramatically over the last few years. RF and microwave therapies for cancer in humans are well documented, and are presently used in many cancer centers. RF treatments for supraventricular arrhythmias, and more recently for ventricular tachycardia (VT) are currently employed by major hospitals. RF/microwaves are also used in human subjects for the treatment of benign prostatic hyperplasia (BPH), and have gained international approval, including approval by the United States Food and Drug Administration (FDA). In the last two years, several otolaryngological centers in the United States have been utilizing RF to treat upper airway obstruction and alleviate sleep apnea. Despite these advances, considerable efforts are being expended on the improvement of such medical device technology. Furthermore, new modalities such as microwave-aided liposuction, microwave assisted tissue anastomoses in future endoscopic surgery, RF/microwave for the enhancement of drug absorption, and microwave septic wound treatment are continually being researched.

18.1 RF/Microwave Interaction with Biological Tissue

Definitions: In this chapter, we detail two types of thermal therapies: RF and microwave. We define RF as frequencies in the range between hundreds of KHz to a few MHz, and microwaves as those in the range of hundreds of MHz to about 10 GHz.

18.1.1 RF Energy [1–7]

The study of the effects of RF current on tissue began in the 1920s, with the work of W.T. Bovie who studied the use of RF for cutting and coagulating. The first practical and commercially available RF lesion generators for neurosurgery were built in the early 1950s by S. Aranow and B.J. Cosman and operated at around 1 MHz [1,2]. The controlled brain lesions made had smooth borders, an immediate improvement over those obtained with DC. As far as the choice of frequency, Alberts et al. have shown that frequencies of up to 250 KHz have stimulating effects on the brain. Thus, RF above 250 KHz was indicated [4–7]. The RF generator is a source of RF voltage between two electrodes. When the generator is connected to the tissue to be ablated, current will flow through the tissue between the active and dispersive electrodes. The active electrode is connected to the tissue volume where the ablation is to be made, and the dispersive electrode is a large-area electrode forcing a reduction in current density in order to prevent tissue heating. The total RF current, I_{RF} is a function of the applied voltage between the electrodes connected to the tissue and the tissue conductance. The heating distribution is a function of the current density. The greatest heating takes place in regions of the highest current density, J. The mechanism for tissue heating in the RF range of hundreds of KHz is primarily ionic. The electrical field produces a driving force on the ions in the tissue electrolytes, causing the ions to vibrate at the frequency of operation. The current density is $J = \sigma E$, where σ is the tissue conductivity. The ionic motion and friction heat the tissue, with a heating power per unit volume equal to J^2/σ. The equilibrium temperature distribution, as a function of distance from the electrode tip, is related to the power deposition, the thermal conductivity of the target tissue, and the heat sink, which is a function of blood circulation. The lesion size is, in turn, a function of the volume temperature. Many theoretical models to determine tissue ablation volume as a function of tissue type are available, but none is as good as actual data.

18.1.2 Microwave Energy

The need for accurate data on permittivity at microwaves and millimeter waves has long been recognized [8] and since 1980 many papers have appeared giving fairly extensive coverage of data collected at frequencies of up to 18 GHz [9–13]. The most recent tabulations of complex permittivity of various biological tissues are reported by Duck [14] and Gabriel et al [15]. Many of the applications and recent advances in the knowledge of the dielectric properties of tissues have been reviewed in the literature [9,12,16,17]. Since 1950, efforts have been directed toward characterization of a variety of tissues at microwave frequencies. Among many reported works in the literature Gabriel et al. [18] provide detailed measurements of a variety of tissues subjected to frequencies up to 20 GHz; they also fit the measured results to a Cole-Cole model with multiple relaxation time constants [19]. Knowledge of the dielectric properties of biological tissues and the basic physical properties of water in tissue at microwave frequencies is essential in order to predict the interaction between field and tissue, which in turn, provides the basis for some of the thermal applications of microwaves described in this chapter.

For sinusoidal fields of frequency f, the permittivity and conductivity are conveniently represented by a single parameter, the complex permittivity ε^*,

$$\varepsilon^* = \varepsilon' - j\varepsilon'' \tag{18.1}$$

where

$$\varepsilon'' = \frac{\sigma}{\omega\varepsilon_0} \tag{18.2}$$

and

$$\omega = 2\pi f \tag{18.3}$$

The real part ε' is referred to as the "relative permittivity" and is related to the energy storage. The imaginary part ε'' is called the "dielectric loss," and corresponds to the power absorption in terms of electromagnetic field interaction with matter; the former influences the phase of the transmitter wave, whereas the latter impacts its amplitude.

When a constant voltage is suddenly impressed on a system initially at equilibrium, for simple systems, the resulting response is usually found to be an exponential function of time. This response, for example, may be the charge buildup at an interface between two different dielectrics, or the alignment of dipoles with an applied electric field. When the voltage is removed, the system relaxes exponentially to its original state. In the general case of an alternating voltage of field, it can be shown [20–22] that the complex permittivity of such simple systems varies with frequency, and can be expressed in the form,

$$\varepsilon^*(\omega) = \varepsilon_\infty + \frac{\varepsilon_S - \varepsilon_\infty}{1 + j\omega\tau} \tag{18.4}$$

where τ is the time constant of the exponential relaxation process, ε_∞ is the permittivity at $\omega \ll 1/\tau$, and ε_S is the permittivity at $\omega \gg 1/\tau$. This is the Debye dispersion equation. The characteristic frequency f_c is defined as,

$$f_c = 1/2\pi\tau \tag{18.5}$$

By separating the Debye equation into real and imaginary parts,

$$\varepsilon' = \varepsilon_\infty + \frac{\varepsilon_S - \varepsilon_\infty}{1 + \left(\frac{f}{f_c}\right)^2} \tag{18.6}$$

$$\varepsilon'' = \frac{\left(\varepsilon_S - \varepsilon_\infty\right) f/f_c}{1 + \left(\frac{f}{f_c}\right)^2} \tag{18.7}$$

or

$$\sigma_d = \frac{\left(\varepsilon_S - \varepsilon_\infty\right) 2\pi\varepsilon_0 f^2}{f_c\left(1 + \left(\frac{f}{f_c}\right)^2\right)} \tag{18.8}$$

where the subscript d denotes the conductivity due to a Debye relaxation process. The measured conductivity, σ_m, will be higher if there are other loss mechanisms in the material, that is,

$$\sigma_m = \sigma_d + \sigma_S \tag{18.9}$$

where σ_s may be the conductivity due to ions and any other contributions at frequencies well below f_c.

If for $f \gg f_c$ we denote σ_d by σ_∞, then

$$\left(\sigma_\infty - \sigma_S\right) = 2\pi f_c \left(\varepsilon_S - \varepsilon_\infty\right)\varepsilon_0 \tag{18.10}$$

which states that the total change in conductivity is proportional to the total change in permittivity and to the characteristic frequency.

The dielectric properties of most materials are not described exactly by the Debye equation. The dielectric properties can then often be approximated empirically by the Cole–Cole equation [19,22]

$$\varepsilon^*(\omega) = \varepsilon_\infty + \frac{\varepsilon_S - \varepsilon_\infty}{(1 + j\omega\tau)^{1-\alpha}} \tag{18.11}$$

It can be shown that this equation is valid for a distribution of relaxation times about a mean value. The Cole–Cole parameter α ranges from 0 to 1 and is an indication of the spread of relaxation times (for $\alpha = 0$, the Cole–Cole equation reduces to the Debye equation).

Depth of penetration (δ) of energy into tissue is defined as the distance in which the power density of a plane wave is reduced by the factor e^{-2}, which numerically is 0.135. Since the power density diminishes as e^{-aL} as the microwave energy travels into the lossy material, the attenuation factor (a) is inversely related to the depth of penetration: ($a = 1/\delta$). The generally accepted value for the depth of penetration of 2450 MHz energy into muscle tissue is 17 mm, and the corresponding attenuation constant is 0.118 for distances in millimeters. For a non-expanding plane wavefront, the reduction of power density as it penetrates muscle tissue is shown in Figure 18.1 [71] (Table 18.1 [70,74]). Figure 18.1 shows, for example, that 11% of the total input power is absorbed in the first millimeter of penetration, and that a total of

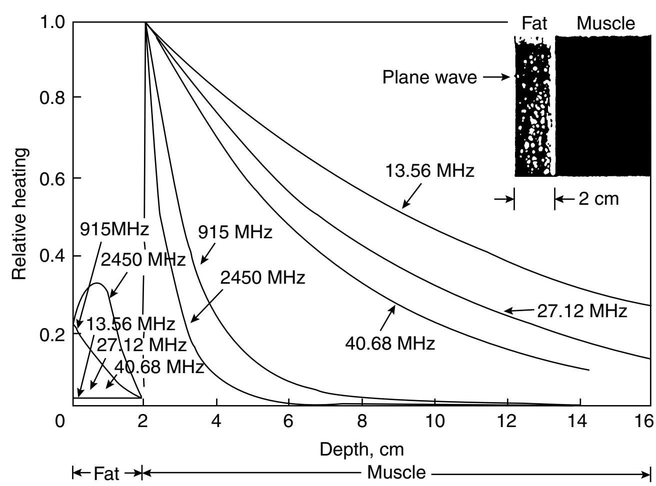


FIGURE 18.1 Calculated relative heating in fat and muscle as a function of distance for five frequencies. (After Paglione.[74])

TABLE 18.1 Relative Permittivity and Conductivity of Biological Media at Microwave Frequencies[70,74]

Frequency (MHz)	Wavelength (cm)	High Water Content Media		Low Water Content Media	
		c	σ (S/m)	c	σ (mS/m)
10	3,000	160	0.625	—	—
100	300	71.7	0.889	7.5	19.1–75.9
300	100	54	1.37	5.7	31.6–107
915	32.8	51	1.60	5.6	55.6–147
2,450	12.2	47	2.21	5.5	96.4–213
3,000	10	46	2.26	5.5	110–234
5,000	6	44	3.92	5.5	162–309
10,000	3	39.9	10.3	4.5	324–549

21% of the input is converted to heat in the first two millimeters of the tissue. Microwave antennas that are utilized in therapeutic medicine, however, operate in the near field. The penetration depth is considerably lower depending on the antenna type) [23].

Previous studies [24] have shown that heat-induced damage to biological tissue is dependent on both the temperature and its duration. The temperature threshold for damage rises as the duration of exposure is shortened.

18.1.3 Test Fixture Structures for Biological Tissue Characterization [25]

A number of techniques have been established for microwave characterization of biological tissues. These techniques are subdivided into TEM transmission lines (i.e., coaxial lines) and non-TEM structures. The majority of published work deals with coaxial transmission lines either as a dielectric loaded [26,27] or an open-ended [28] coaxial line. In both approaches, the change in the terminating impedance causes change in the input reflection coefficient of the line. A different complex permittivity of the tissue under test causes change in the capacitance and conductance of termination, hence impacting amplitude and phase of the reflected wave. This technique is popular and has been developed extensively by Burdette et al. [28] and Stuchly et al [29].

Dielectric loaded waveguide structures such as circular or rectangular cross-section metallic waveguides are the second most popular structures for tissue characterization. Steel et al. [30,31] reported a technique for characterization of liquids and solids. For liquids measurements the sample is contained in a length of waveguide and a moving short circuit enables the liquid thickness to be varied. A microwave signal is applied to the sample, the modulus of the reflected signal is recorded as a function of sample length, and a least-squares curve-fitting analysis of data enables various parameters to be obtained. For measuring solid samples, an automated slotted line is used to record the standing wave ratio in front of the sample, the latter terminated by a short circuit.

The above techniques all use the reflecting wave from a terminating impedance and the transmission line theory to extract the electrical parameters of the tissues under test. On the other hand, methods exist that use tissue samples to change the resonance frequency and quality factor of a cavity resonator [32] More accurate results can be achieved by using simpler setups. However, the resonance methods are only applicable to discrete frequency points corresponding to the resonance frequencies of the modes of interest. For instance, Land and Campbell [32] present a technique that uses simple formulas that relate the complex permittivity of a small piece of tissue to the change in resonance frequency and quality factor of a cylindrical cavity. The cavity is filled with PTFE and resonates for TM_{010} mode; it has a diameter of 50.8 mm and a length of 7.6 mm. Three 1.5 mm diameter sample holes are provided in the cavity. The cavity resonates at 3.2 GHz. The microwave setup is very simple, that is, a signal generator, frequency meter, directional coupler, attenuator, diode detector, and a voltmeter are used in the experiment.

18.1.4 Tissue Characterization through Reflection Measurements [25]

In order to characterize various biological tissues, three distinctive methods have been reported. The first method is based on voltage standing wave ratio measurements using slotted line waveguides [31]. The second approach is based on an impedance analyzer and is suitable for microwave frequencies.

Finally, the most popular approach in the last 20 years has been the use of a network analyzer [28,33–38]. With the advent of the automatic network analyzer (ANA) (e.g., Hewlett-Packard's HP8510, HP8753, HP8720), accurate measurements of scattering parameters have been extended to frequencies beyond 10 GHz (since 1984). The advantage of this technique is fast and accurate measurements of coaxial-based structures. The majority of attempts to accurately characterize the complex permittivity of biological tissues in the last 10 years are based on these families of ANAs [18,38,39].

18.1.5 Microwave Antenna in Therapeutic Medicine: Issues

Biomedical antenna designs have typically addressed the applications of microwave hyperthermia for the treatment of malignant tumors, microwave catheter ablation for the treatment of cardiac arrhythmia, microwave balloon angioplasty, and microwave-assisted liposuction. The analytical basis for much of this work has been based on the lossy transmission line analysis developed by King et al., [40,41] which allowed the calculation of input impedance and near fields for simple antenna geometries. Several researchers have refined this work to provide improved accuracy or wider applicability. For example, Iskander and Tumeh developed an iterative approach to designing multi-sectional antennas based on an improved King method [42] and have used this approach to compare different antennas [43]. Debicki and Astrahan developed correction factors to allow accurate modeling of the input impedance for electrically small multi-section antennas, [44] and Su and Wu have refined the King approach to determine the input impedance characteristics of coaxial slot antennas [45]. Casey and Bansal used a different approach than King to compute near fields of an insulated antenna using direct numerical integration of a surface integral [46].

A problem inherent in many biomedical antenna designs is the effect of heating along the transmission line due to current flow on the outer conductor of the coaxial transmission line. In most applications this effect is undesirable since thermal energy is being delivered to healthy tissue outside the intended treatment area. Moreover, the magnitude of this effect is a strong function of the insertion depth of the antenna. Similarly, the antenna input impedance also varies with the insertion depth. Hurter et al. [67] proposed the use of a sleeve Balun to present a high impedance to the current on the outer conductor, thus concentrating the microwave energy at the antenna tip. Temperature profile measurements made in a phantom using a fiber-optic temperature probe clearly show improved localization of thermal energy delivery.

18.2 RF/Microwaves in Therapeutic Medicine

18.2.1 RF/Microwave Ablation for the Treatment of Cardiac Arrhythmias [47,48]

Cardiac arrhythmias can result from a variety of clinical conditions, but at their root is an abnormal focus, or pathway, of electrical activity. Abnormal sources of electrical activity most commonly occur at or above the AV-node, and are thus deemed supraventricular tachy-arrhythmias. Alternatively, abnormal ventricular foci cause ventricular tachycardia. The presence of abnormal conduction pathways can also result in an uncontrolled cycling of electrical activity resulting from retrograde signal conduction through the myocardium (reentry tachyarrhythmias). Reentry can occur within the AV-node (AVNRT), or via accessory conduction pathways (AP). Regardless of the specific etiology, once the source of the arrhythmia has been identified, destruction of the abnormal cardiac tissue is curative. The goal of ablation is to modify the electrical system of the heart by converting electrically active cardiac tissue to electrically inactive scar tissue. The scar or lesion that forms then blocks the focus or accessory pathway and prevents the tachycardia. Various energy forms have been used to create such localized tissue injury, including direct current (DC), radiofrequency (RF), and microwave energy (Table 18.2) [47,48].

The clinical use of DC ablation dates back to 1982. An electrode catheter is placed at the desired location, and a DC shock is applied. Although complete ablation occurs in up to 65% of patients, DC

TABLE 18.2 Energy Sources for Catheter Ablation [47]

	Direct Current	Radiofrequency	Microwave
Waveform	Monophasic, damped sinusoidal	Continuous unmodulated sinusoidal	N/A[a]
Frequency	DC	550–750 kHz	915, 2450 MHz
Voltage V	2000–3000 V	<100 V	N/A
Mechanism of injury	Passive heating, baro trauma, electric field effects	Resistive heating	Radiant heating
Sparking, barotrauma	Yes	No	No
General anesthesia	Yes	No	No
Lesion size	Moderate	Small	Unknown
Control of injury	Low	High	High

[a] N/A = data not available.

ablation is fraught with complications. Hypotension, perforation, cardiac tamponade, embolization, pericarditis, and ventricular tachyarrhythmias have been reported in as many as 10% of patients. Mortality associated with DC ablation may be as high as 5% in some patient groups. RF ablation was developed with the hope of decreasing the risks associated with DC application. In RF ablation, lesion formation results from resistive tissue heating at the point of contact with the RF electrode (Figure 18.2). This heating is thought to lead to coagulation necrosis and permanent tissue damage. If there is poor tissue contact, RF current cannot be coupled to the underlying tissue, and the desired effect of tissue heating is lost. Overall success rates for RF ablation have been reported to be as high as 90% for AV junction ablation, and as high as 95% when applied to re-entry-mediated tachycardia. Furthermore, RF ablation has not been reported to result in serious side effects.

The first supraventricular tachycardias targeted for RF ablation were those associated with the Wolff–Parkinson–White syndrome. In this condition the anatomic basis for supraventricular tachycardias is an accessory connection or pathway that connects the atrium and ventricle outside the normal AV conduction pathway (Figure 18.3). These accessory pathways cross between the atrium and the ventricle at the

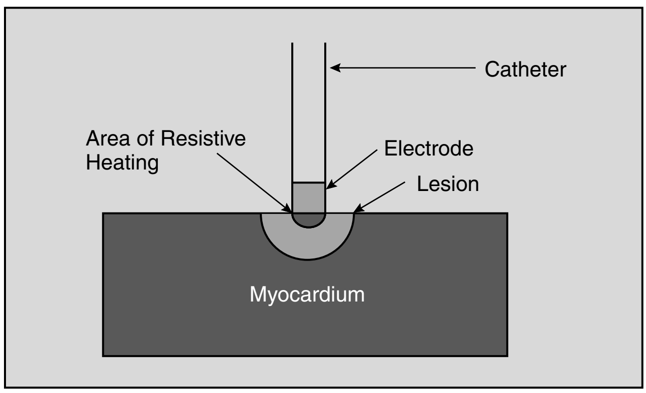


FIGURE 18.2 Mechanism of RF ablation. When RF current is delivered to the tip of a catheter electrode, resistive heating occurs along a small rim of tissue in direct contact with the electrode. A lesion is created as heat conducts passively away from this zone and the surrounding myocardium is heated to a temperature where cell death occurs (~50°C). Lesion size is therefore a function of the size of the electrode and the resulting temperature at the electrode-tissue interface.

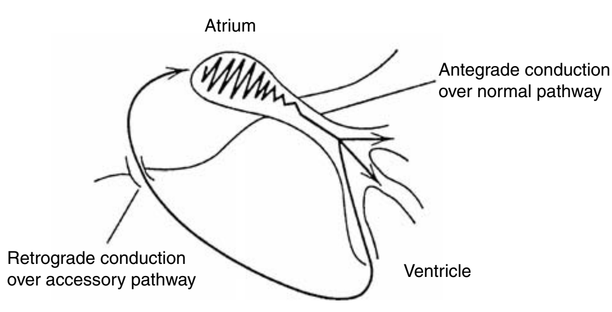


FIGURE 18.3 An arrhythmic circuit associated with the Wolff–Parkinson–White syndrome. In this syndrome, there is a connection between the atrium and ventricle outside the normal V nodal pathway (accessory pathway). A tachycardia circuit can develop if an impulse conducts antegrade (forward) via the normal V node pathway and is able to conduct retrograde (reverse) from the ventricle to the atrium via the accessory pathway. Catheter ablation successfully treats these arrhythmias because it interrupts accessory pathway conduction without interfering with normal AV nodal conduction.

level of the mitral and tricuspid anulus. RF energy delivered to the mitral or tricuspid anulus either from the ventricular or atrial aspect can ablate these pathways (Figure 18.4). Success rates of over 90% have been reported using either approach.

Recently, encouraging results in the treatment of ventricular arrhythmias occurring as a consequence of diffuse processes such as myocardial ischemia or infarction have been published. The search for ablation modalities capable of safely generating even larger lesions has spawned an interest in microwave ablation (Figure 18.5). Unlike DC and RF techniques, which generate lesions of relatively limited size and penetration,

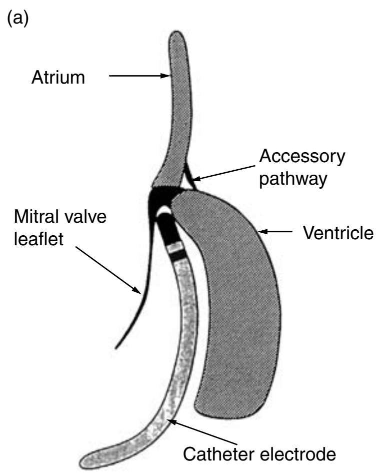


FIGURE 18.4 Diagrams of electrode positions used in RF catheter ablation of accessory pathways: (a) for the ventricular approach a catheter is passed retrograde across the aortic valve and positioned under the mitral leaflet; (b) for the atrial approach a catheter is passed across the interatrial septum (trans-septal catheterization) and positioned on top of the mitral valve leaflet. Electrical mapping confirms the site of the accessory pathway prior to the delivery of RF energy.

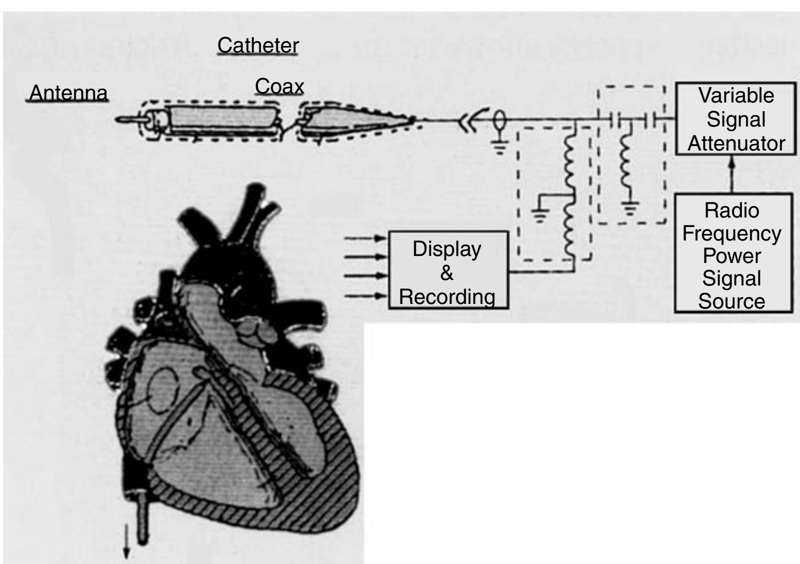


FIGURE 18.5 Microwave system used for myocardial tissue ablation.

microwave energy might allow for greater tissue penetration, and thus a greater volume of heating. Microwave ablation systems are currently being developed.

Microwave hyperthermia has been useful in radiation oncology for the treatment of various solid tumors [49,50,51]. The cardiac applications of this modality have only recently been explored. Microwave energy using either 915 or 2450 MHz has been studied in an attempt to enlarge myocardial lesions in catheter ablation [52,53,54]. Microwave energy is delivered down the length of a coaxial cable that terminates in an antenna capable of radiating the energy into tissue. Radiant energy will cause the water molecules in myocardial tissue to oscillate, producing tissue heating and cell death. The higher frequency of microwave energy allows for greater tissue penetration and, theoretically, a greater volume of heating than that possible with RF, which produces direct ohmic or resistive heating.

Wonnell and coworkers studied the effects of microwave energy for cardiac ablation using a helical antenna mounted on a coaxial cable (2.44 mm o.d.) [55]. High-frequency current at 2450 MHz was delivered via the helical antenna into a tissue-equivalent phantom model. The temperature distribution profile was measured around the antenna as well as into surrounding volume (the depth of penetration). The volume of heating for the microwave catheter system was 11 times greater than that of an RF electrode catheter at the same surface temperature. In addition, the microwave catheter penetrated an area that was twice as large as that penetrated by the RF catheter. These data suggest that microwave energy will produce larger lesions than those produced by RF because a greater volume of tissue is being heated. An additional theoretical advantage of the microwave system is that direct tissue contact is not crucial for tissue heating since heating occurs via radiation, and not via direct ohmic heating as seen with RF. Using this system, preliminary studies in six animals demonstrated that complete heart block could be achieved in all six animals by directing microwave energy (50 W at 2450 GHz for about 200 seconds) to the atrioventricular junction.

We evaluated helical and whip antenna designs in a tissue-equivalent phantom at 915 MHz and 2450 MHz utilizing a coaxial cable (0.06 in o.d.) [56]. All catheters were measured utilizing a network analyzer prior to placing them in the phantom model. Such analysis demonstrated the great variability in tuning of these microwave catheters. Microwave ablation catheters have suffered from imperfect tuning leading to inefficient radiation of energy. Consequently, there is generation of heat along the length of the catheter rather than radiation of energy into tissue. Little heating into tissue was observed in poorly tuned catheters. Such analysis underscores the critical importance of proper tuning of microwave catheters prior to any further studies.

A perfusion chamber containing a muscle-equivalent phantom was constructed and placed in a saline bath held at 37°C. The muscle-equivalent phantom consisted of TX150, polyethylene powder, NaCl, and water. Ablation catheters were placed on the surface of the phantom material. Temperature measurements were performed using a 12-channel Luxtron fiber-optic thermometry system. Probes were placed beneath the surface of the phantom. Saline at a constant temperature of 37°C was infused at a flow rate of 4 L/min across the surface of the phantom. This model simulates the heart where the phantom material has the dielectric properties of cardiac muscle and the saline properties of blood (Figure 18.6).

Temperature curves were plotted from probes placed 1, 2.5, 5, and 7.5 mm from the point of maximal heating on the microwave catheter. Thermal profiling of these catheters demonstrated volume heating. Heating was proportional to power duration and to surface temperature. In addition to the volume heating, conductive heating was also present as a result of the increased temperature at the catheter-phantom interface [56].

In vivo ablation using microwaves was performed on canine left ventricular myocardiums. A power of 80 W was delivered for a total of 5 min. Mean lesion size measured 435 ± 236 mm^3, which was similar in size to lesions created with small-tipped RF catheters. The microwave ablation catheters, as presently designed, were not capable of producing lesions larger than those produced by RF catheters [56].

Practical problems remain to be solved before microwaves become a useful clinical energy source. These problems include (1) power loss in the coaxial cable, (2) resultant heating of the coaxial cable during power delivery that has led to a breakdown in the dielectric and catheter material, (3) inefficiency of the radiating antenna, and (4) lack of a unidirectional antenna that can radiate energy into tissue and

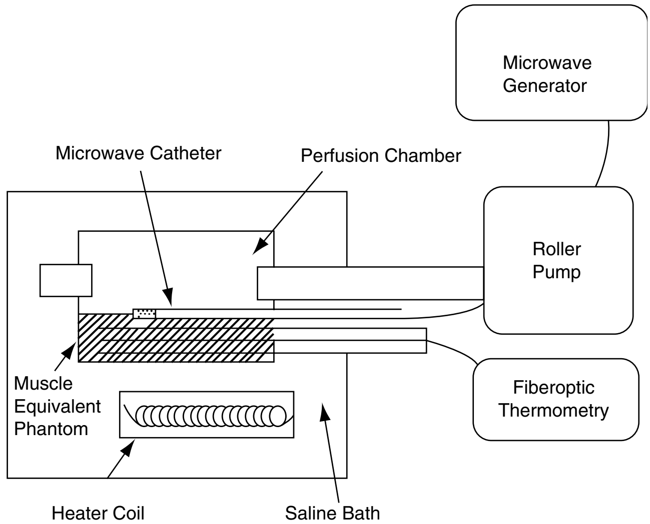


FIGURE 18.6 Flow-phantom model for cardiac ablation catheters.

not into the circulating blood pool. At the present time microwave catheter systems are poorly efficient radiators of energy into cardiac tissue. These obstacles will have to be overcome before microwaves supplant radio frequency as the preferred energy source for cardiac ablation.

18.2.2 RF/Microwave Treatment of BPH [57]

Benign prostatic hypertrophy (BPH) is an enlargement of the prostate gland that can lead to compression of the urethra, and thus cause urinary tract obstruction. The prostate gland is an organ at the base of the male bladder that surrounds the urethra and produces seminal fluid. Overgrowth of prostatic tissue leads to compression of the urethra. BPH is among the most common medical conditions affecting men over the age of 50. In fact, over 50% of men over 50 years of age have enlarged prostates. Symptoms of urinary tract obstruction (frequent urination, decreased urine flow, nocturia, dribbling, discomfort, pain) manifest themselves most commonly at 65–70 years of age.

Although drug therapy may be effective for patients with early stages of BPH, many men will need invasive intervention for relief of symptoms. Surgical excision of prostatic tissue has been the standard care for more advanced forms of BPH. Procedures such as prostatectomy and transurethral resection of the prostate, however, carry significant risks. To minimize hazards such as hemorrhage, coagulopathies, pulmonary emboli, bladder perforation, incontinence, infection, urethral stricture, retention of prostatic chips, infertility, and retrograde ejaculation, minimally invasive alternatives have been developed and are being investigated. Transurethral RF and microwave procedures are becoming promising alternatives to surgical intervention. The goal of therapy is to decrease the volume of prostatic tissue. RF Transurethral Needle Ablation (TUNA) (Figure 18.7) involves the introduction of interstitial needle electrodes directly into prostatic tissue. This technology uses RF (460 KHz) with excellent control of the RF thermal energy applied to the tissue. The TUNA catheter used is 24.1 cm long and 21 French. Through the tip of the catheter, two needles (electrodes) oriented 40° apart, can be deployed. The electrode-needles are shaped to facilitate passage through tissue. They are thin, and thus can be directed from the catheter through intervening tissue with a minimum of trauma to normal tissue. Each electrode-needle is enclosed within a longitudinally adjustable sleeve acting as a shield to prevent exposure of the tissue adjacent to the sleeve to the RF current, thus preserving the urethra by reducing the possibility of a rise in its temperature. The sleeve is also used to control the tissue interface, and therefore the ablation volume. Both the electrode-needle and the sleeve are locked into position. The TUNA catheter needle acts as the thermal

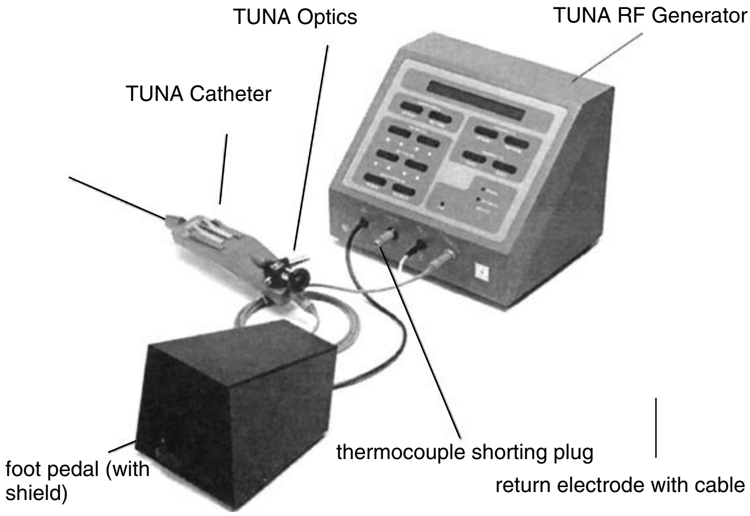


FIGURE 18.7 TUNA RF generator unit (with permission of VidaMed, Inc.).

electrode, and a grounding pad that is placed in back of the subject under treatment closes the RF circuit to the power supply (Figure 18.8).

Thermocouples are located at the shield tip below each needle, and at the catheter bullet head (in order to record ablation temperature and prostatic urethral temperature, respectively). The RF unit (VidaMed, Inc.) includes an RF generator with the following readouts: RF power level, ablation time, impedance, and six thermocouple readouts (Figure 18.7). The TUNA catheter (Figure 18.8a) includes direct fiber-optic vision, as well as provisions for introducing electrode-needles at various angles (Figure 18.8b).

Transurethral Microwave Thermotherapy (TUMT) (Figure 18.9a–c) has also shown promise as a therapeutic modality for the treatment of BPH. This technique uses a microwave delivery system housed within a transurethral catheter. Its goal is to destroy prostatic tissue selectively without damaging the urethral mucosa or structures surrounding the treatment area. At microwave frequencies, temperatures in the target tissue can be raised to as high as 45–70°C without damaging periprostatic tissue. TUMT is used routinely worldwide.

18.2.3 Microwave Balloon Catheter

18.2.3.1 Microwave Balloon Angioplasty [58,59]

Atherosclerosis, with its resultant occlusion of coronary blood flow, remains a leading cause of morbidity and mortality. For many patients with advanced disease, or in whom pharmacologic management has failed, percutaneous transluminal balloon angioplasty (PTCA) has offered an effective alternative to coronary bypass surgery. The efficacy of PTCA, however, has been limited by restenosis rates ranging from 17% to 47%, as well as by a risk of arterial dissection and/or thrombus formation. Furthermore, acute occlusion, resulting from elastic recoil at the angioplasty site, can occur in as many as 5% of patients undergoing PTCA. Such patients require emergency heart surgery. Microwave balloon angioplasty (MBA), the first microwave application in cardiology, was developed with the ultimate goal of decreasing both acute and long-term restenosis risks.

MBA, like PTCA, employs a balloon catheter that is advanced to the site of arterial stenosis. While PTCA uses only the pressure generated by balloon inflation to dilate the affected artery, MBA takes advantage of the volume heating properties of microwave emitters. In MBA, a microwave cable-antenna assembly is threaded through the catheter, with the antenna centered in the balloon portion of the catheter

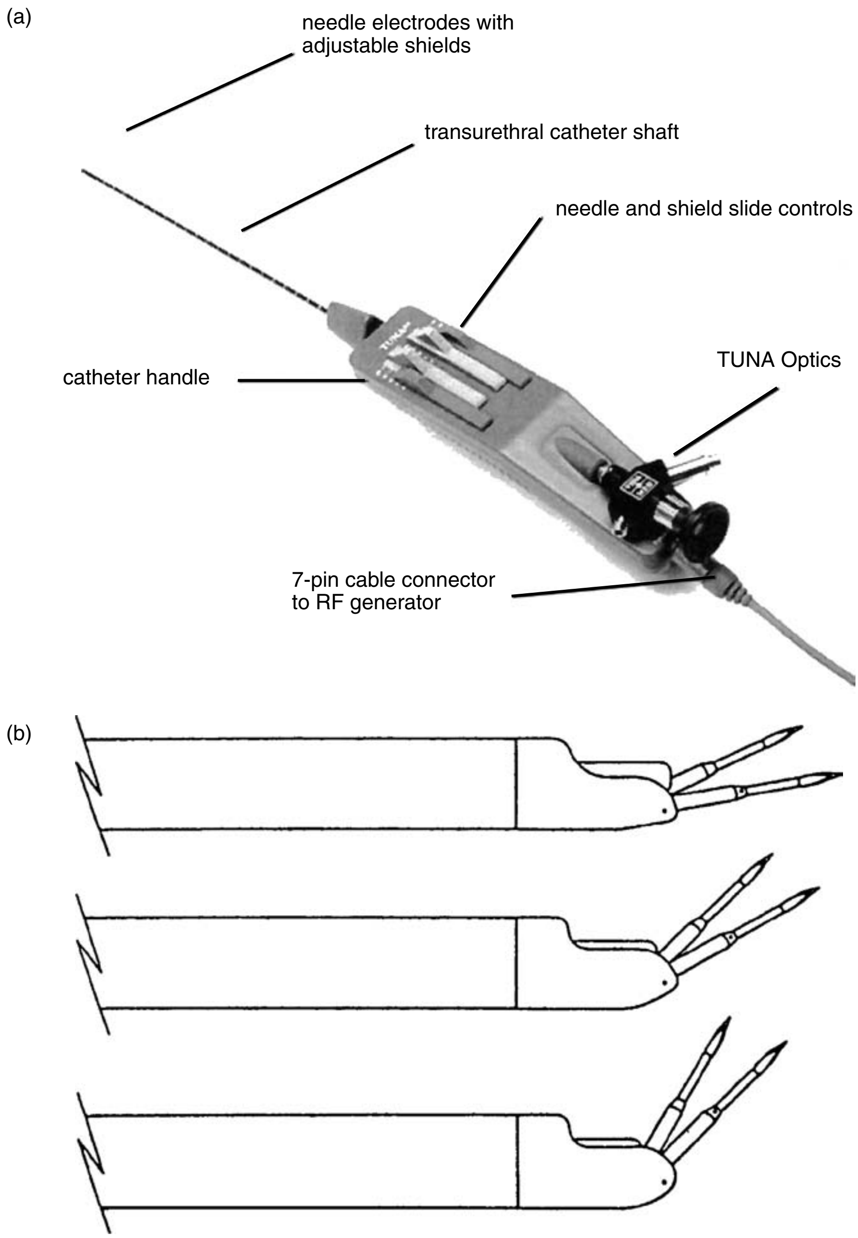


FIGURE 18.8 (a) TUNA catheter with handle incorporating direct fiber-optic vision (With permission of VidaMed, Inc.). (b) Electrodes and needles at various angles. (With permission of VidaMed, Inc.).

(Figure 18.10). By heating the tissue as the balloon is inflated, it was hoped that a patent vessel would be created that would be resistant to both acute and chronic reocclusion. Early *in vivo* studies, at 2.45 GHz, were conducted to assess the effects of various energy levels upon normal and atherosclerotic rabbit iliac arteries. Research on the therapeutic potential was subsequently conducted on atherosclerotic rabbit iliac arteries using microwave energy to raise the balloon surface temperature to 70–85°C. When compared to simultaneously performed conventional angioplasty, MBA at 85°C produced significantly wider luminal diameters, both immediately after angioplasty and 4 weeks after the procedure (Figure 18.11).

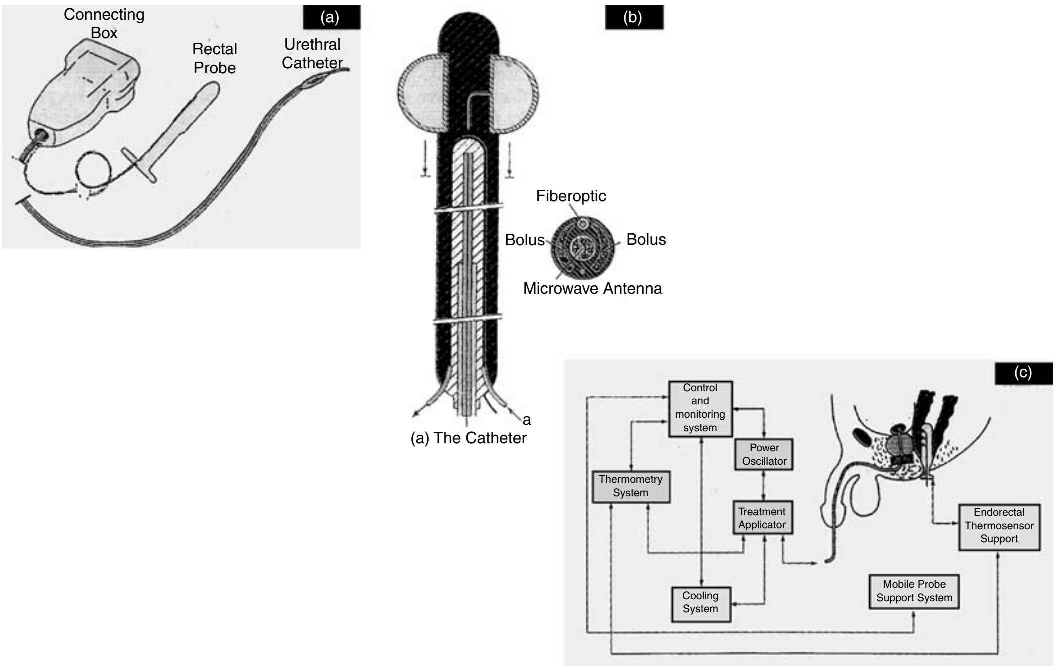


FIGURE 18.9 (a) Schematic representation of the treatment catheter; (b) cutaway view; (c) Prostatron treatment functional diagram. (With permission from Technomed Medical Systems.)

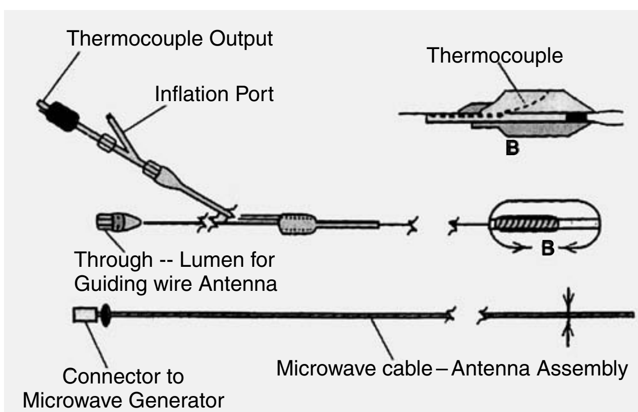


FIGURE 18.10 Microwave balloon angioplasty system.

Further work, utilizing mongrel dogs with thrombin-induced coronary occlusion, has demonstrated the feasibility of MBA as a treatment modality for coronary thrombosis. MBA of such coronary thrombi in dogs resulted in patent vasculature with the added benefit of an organized and stabilized thrombus. Although the technique described was successful in animal studies, it has not yet found its way into clinical use. However, microwave balloon angioplasty has recently been suggested for applications in carotid stenosis and occlusions in peripheral circulation.

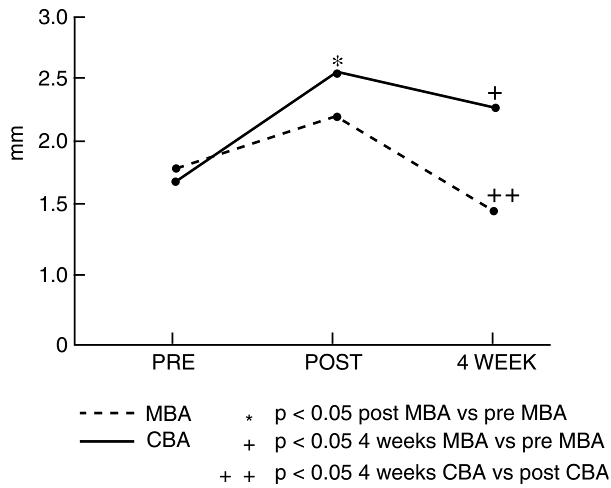


FIGURE 18.11 Microwave thermal angioplasty (85°C).

18.2.3.2 Microwave Balloon Catheters in the Treatment of Benign Prostatic Hypertrophy [60]

Localized microwave hyperthermia has been used for more than a decade to treat cancer of the prostate and since 1985 to treat BPH. The initial hyperthermia treatments used microwave applicators that heated the prostate via the rectum, but today transurethral applicators are favored. Transurethral applicators are usually placed inside liquid-cooled catheters and the temperatures produced inside the treated prostate can be measured noninvasively with a microwave radiometer.

With balloon catheters it is possible to produce both high therapeutic temperatures throughout the prostate gland without causing tissue burning, and biological stents in the urethra in a single treatment session. Compared to conventional microwave catheters, the distances microwaves have to travel through the prostate to reach the outer surface of the gland are reduced by the use of balloon catheters, as is the radial spreading of the microwave energy (Figure 18.12). Furthermore, compression of the gland tissues reduces blood flow and its cooling effect within the gland. Also, since catheter balloons make excellent contact with the urethra, much better than do conventional catheters, the urethra is well cooled by the cooling liquid within the balloon, and is therefore well protected from thermal damage.

18.2.3.3 Microwave Balloon Catheters in the Treatment of Cancer [60]

Interstitial hyperthermia is usually combined with radiation therapy using radioactive seeds that are inserted into the tumor via the same tubing that is used for the hyperthermia (Figure 18.13). A typical treatment sequence is brachytherapy (irradiation of the tumor with the seeds) followed by hyperthermia, followed again by brachytherapy. The interstitial hyperthermia enhances the efficacy of brachytherapy because (1) hyperthermia interferes with the repair of cells that have been sublethally damaged by the ionizing radiation, (2) cells in the S phase of the cell cycle and hypoxic tumor cells tend to be resistant to ionizing radiation but sensitive to heat, and (3) hyperthermia can be effective in oxygenating radiation-resistant hypoxic cells.

Intersitial arrays using conventional applicators are useful only for treating small tumors, because each interstitial applicator can heat and irradiate only a small volume of tissue, and the number of applicators that can be inserted into a tumor is limited because of their invasive nature. Interstitial applicators using balloons, on the other hand, can heat much larger volumes of tissues than can conventional interstitial applicators, making it possible to treat larger tumors. A catheter with a deflated balloon at its tip is inserted into the tumor volume to be heated, or where applicable, into a natural opening of the body

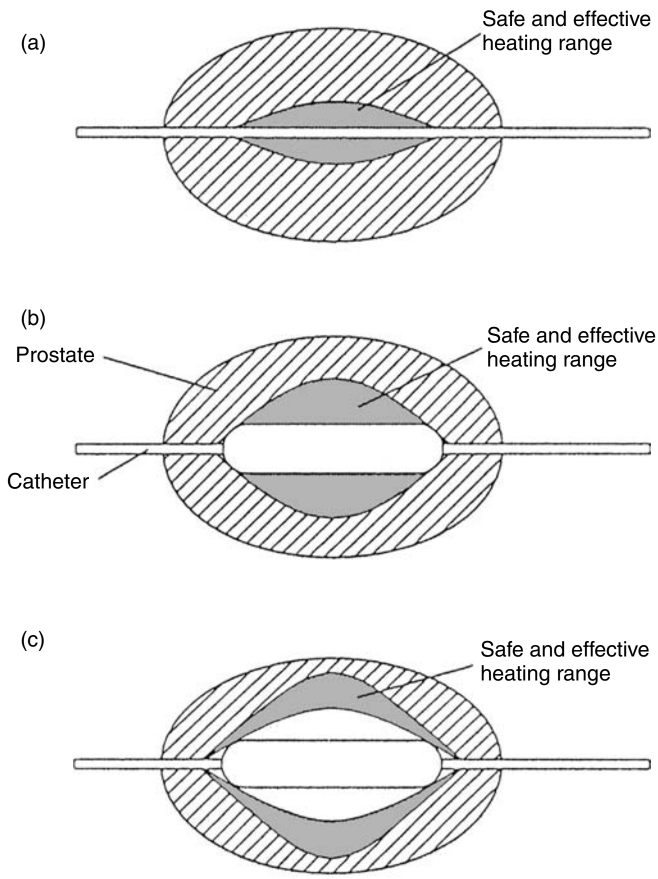


FIGURE 18.12 Safe and effective heating ranges in prostate glands with regular microwave catheters and with microwave balloon catheters. Microwave heating patterns: (a) with regular catheter; (b) with balloon catheter; (c) with balloon catheter and water cooling. (After Sterzer [60].)

such as the urethra, rectum, or vagina; the balloon is inflated, and radioactive seeds or a microwave antenna are inserted through the center lumen of the balloon catheter (Figure 18.14).

18.2.4 RF in the Treatment of Obstructive Sleep Apnea [61–63]

Obstructive sleep apnea (OSA) is a disorder diagnosed when an individual's upper airway becomes intermittently blocked during sleep and breathing becomes interrupted. Approximately 20 million Americans are estimated to suffer from OSA, and over half of these are between the ages of 30 and 60 years. During sleep, there is a relaxation of the structures surrounding the pharynx/throat. Breathing becomes interrupted (apnea) when these anatomical structures relax in a position that occludes airflow. The most commonly involved structures include the soft palate, the base of the tongue, and the tonsils/adenoids. Enlarged turbinates within the nose can serve to further impede airflow.

OSA and its resultant interruption of normal sleep patterns have a wide range of clinical effects. Patients may experience daytime sleepiness, most hazardous while driving or during work. They may also exhibit personality changes, difficulty in concentrating, memory difficulties, headaches, or sexual dysfunction. Sleep apnea is also associated with increased rates of systemic and pulmonary hypertension, stroke, heart failure, and myocardial infarction.

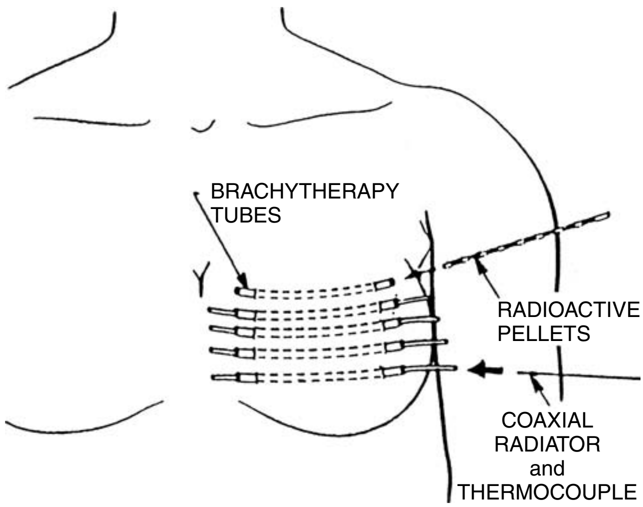


FIGURE 18.13 Interstitial hyperthermia combined with brachytherapy for treating breast cancer. (After Sterzer [60].)

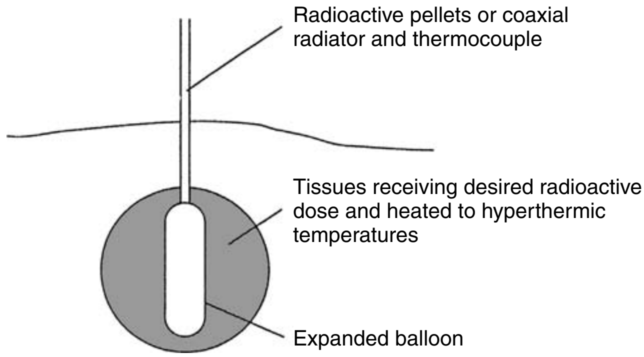


FIGURE 18.14 Interstitial microwave balloon catheter for combined localized hyperthermia and brachytherapy treatment of cancer. (After Sterzer [60].)

Treatment depends on the severity and frequency of symptoms. Some mild cases may be managed with weight loss alone. Often, however, further intervention is needed. Conventional management has relied upon dental appliances to maintain an open airway, ventilators to provide Continuous Positive Airway Pressure (CPAP), and attempts at surgical correction of the airway obstruction. Though effective, dental appliances and CPAP are both uncomfortable, and suffer from relatively low patient compliance rates (40–70%). Surgical correction may involve either excision of "excess tissue" (uvulopalatopharyngoplasty) or more involved maxillofacial surgery. Surgical cure rates have been reported to range between 30% and 75%.

Recently, Somnus Medical Technologies has developed an RF system (Somnoplasty™) that uses needle electrodes to create precise regions of submucosal tissue coagulation. Thus, both the tissue volume and the resulting airway obstruction are reduced. Applicator probes have been developed to target specific tissues including the base of the tongue (Figure 18.15a), the uvula (Figure 18.15b) the soft palate

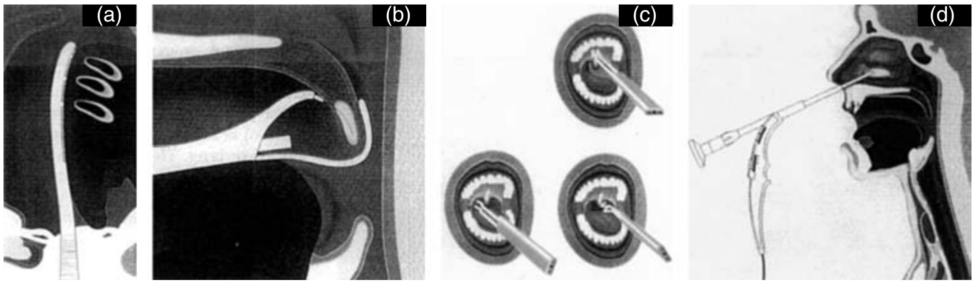


FIGURE 18.15 (a) Tongue Somnoplasty; (b) Uvula Somnoplasty; (c) Palatal Somnoplasty; (d) Turbinate Somnoplasty (With permission of Somnus, Inc.).

(Figure 18.15c), and the nasal turbinates (Figure 18.15d). Somnoplasty is designed to be performed on an outpatient basis, under local anesthesia, and is expected to boast such benefits as immediate results, little postoperative edema or discomfort, and no permanent scarring.

In the paper entitled "Radiofrequency Volumetric Reduction of the Tongue—A Porcine Pilot Study for the Treatment of Obstructive Sleep Apnea Syndrome," Powell et al., [64] reported on the use of RF for the volumetric reduction of the tongue. Powell's three-stage pilot study investigated both the *in vitro* and *in vivo* effects of RF, delivered via a customized needle electrode. Volumetric measurements were performed using implanted ultrasonic crystals positioned around the treatment site. Changes in tissue volume could then be assessed both before and after the delivery of RF energy. To establish the feasibility of the technique two bovine tongues (*in vitro*), were used in the initial stage of the project. A single 0.05-in. diameter needle electrode delivered 30 kj over a 20 minute period at two sites per tongue. Volume reductions of between 12.8% and 26.7% were noted immediately after the procedure, with an additional 4% reduction noted after 4 hours. The second stage was conducted using pigs, *in vivo*, and demonstrated that volume reduction increases as the amount of energy delivered is increased from 6.8 to 40 kj. Finally, in the third stage, an *in vivo* porcine model was again used, this time assessing clinical efficacy of the procedure by measuring both tissue volume changes, and histological changes. RF tissue reduction was performed on 9 pigs, with 3 additional pigs serving as controls. An 0.035-in. diameter needle electrode was used to deliver 2.4 kj over 6 ± 1.20 minutes. Immediately after the procedure, a mean volume shrinkage of 7.02% was described. By 24 hours after the procedure, edema resulted in a 4–6% increase in tissue volume, thus returning nearly to baseline volumes. Subsequently, however, a progressive volume reduction of up to 26.3% was identified over the following 10 days. Animals were sacrificed at between 1 hour and 5 weeks after the procedure. Lesions were described as spherical, well-defined regions of tissue destruction, initially demonstrating edema and hemorrhage. As the lesion healed, scar formation occurred along with neovascularization. Tissue and vessels surrounding the lesion remained intact and viable. Based on the apparent success of this technique in the animal model, RF tissue reduction may offer a promising alternative to the conventional management of obstructive sleep apnea.

18.2.5 Microwave-Aided Liposuction [59,65]

Liposuction is used for aesthetic and reconstructive surgery. Its uses include the undermining of large flaps while preserving vascular attachments, removing lipomas, treating gynecomastia, and reducing axillary hyperhidrosis. The application of microwave for aided liposuction may reduce some problems associated with standard mechanical liposuction, including blood loss, fluid shifts, and systemic effects.

Dry-technique liposuction versus microwave-aided dry-technique liposuction—Preliminary work has been conducted, in swine, to compare the effects of the dry-technique liposuction versus microwave-aided dry-technique liposuction. The "non-microwave" dry-technique liposuction performed at the two cephalad sites yielded typical fat debris that grossly appeared to be mixed with a noticeable amount of blood. The "microwave" liposuction performed at the two caudal sites yielded fat that differed considerably in quality and texture from tissue extracted using the dry-technique. The duration of microwave-aided suctioning appeared to be related to the histologic changes observed in the subcutaneous fat derived from the caudal sites. The fat initially removed during the first 30 s grossly appeared similar to conventionally suctioned fat. However, the fat removed as the duration of microwave-aided liposuction increased from 30 s to 2 min appeared increasingly softened. The longest duration of microwave suctioning, from 2 to 4 min, yielded fat that grossly appeared to be fused into an opaque, amorphous melted state.

The tumescent technique for liposuction surgery—The "tumescent technique" of liposuction was introduced in 1986. Use of the Klein needle has allowed the anesthetic solution to be rapidly injected through the same incision used for liposuction, efficiently anesthetizing large subcutaneous areas, thus eliminating the need and risks of general anesthesia. Injection of a large volume of dilute lidocaine produces a swelling and firmness of the site to be aspirated, which greatly facilitates fat removal. The small (3–4 mm) cannulas produce less trauma and therefore result in less blood loss, bruising, and discomfort. The basic technique was later expanded, and much larger volumes of lidocaine were administered, resulting in the capability of aspirating significantly greater volumes of tissue with a minimum increase in blood loss. This was achieved with serum lidocaine levels well below the toxicity range. We believe that using microwave volume heating will further enhance and benefit the tumescent technique.

Tumescent-technique liposuction versus microwave-aided tumescent-liposuction—A similar protocol was followed at corresponding sites on the left side of the swine. The only modification was to employ tumescent liposuction instead of the dry technique that was used for sites on the right side. The solution used for tumescence consisted of l000 cc of normal saline, combined with 60 cc of 1% lidocaine with epinephrine. Approximately 250 cc of this solution was infiltrated into each of the four sites prior to liposuction. The conventional "non-microwave" tumescent liposuction performed at the two cephalad sites yielded fat typically seen in such procedures; there was also less bleeding than seen with the dry technique.

Tumescent liposuction combined with microwaves between 30 and 40 watts yielded a transformation in the fat suctioned, enabling easier fat removal with less bleeding in comparison to both conventional dry and tumescent liposuction without microwaves.

Cannula design—The cannula utilized was a Byron Accelerator III type cannula, which was modified to hold a microwave semirigid coaxial cable having a whip antenna at the distal end (Figure 18.16). The tip of each cannula distal of the suction port was modified by removal of its metal tip, which was replaced with a dome made of plastic in order to facilitate microwave radiation. The suction port in the proximal end of the cannula handle was converted to accept the semirigid coaxial cable/antenna structure. Suction was effectuated through a new port installed in the cannula handle.

The system used in our preliminary experiments was designed for use at 2.45 GHz while immersed in a tissue phantom. With the modified liposuction cannula and antenna, we have measured return losses as low as –37 db.

Antenna considerations [**66**]—The antenna design for microwave aided liposuction (MAL) presents some interesting challenges. The antenna must deliver microwave energy to heat the treated volume of fat. Unlike traditional biological antenna designs, the MAL antenna radiates in close proximity to a metallic cannula. The cannula serves two primary functions. First, the openings in the end of the cannula are sharp to facilitate fat removal via mechanical cutting. Second, removed fat is suctioned through the cannula. Clearly, the cannula imposes a more complex antenna geometry than exists for traditional biological antennas. Furthermore, as opposed to microwave hyperthermia application, for example, it may actually be desirable to allow heating to occur along the transmission line since that heating will prevent the coagulation of fat being suctioned through the cannula. Analytical antenna design approaches become increasingly difficult under these design constraints. Therefore, it is appealing to consider the use of accurate simulation tools to design the antenna and to analyze the antenna performance. This approach was used by Labonte et al. who implemented a finite-element method (FEM) in the frequency

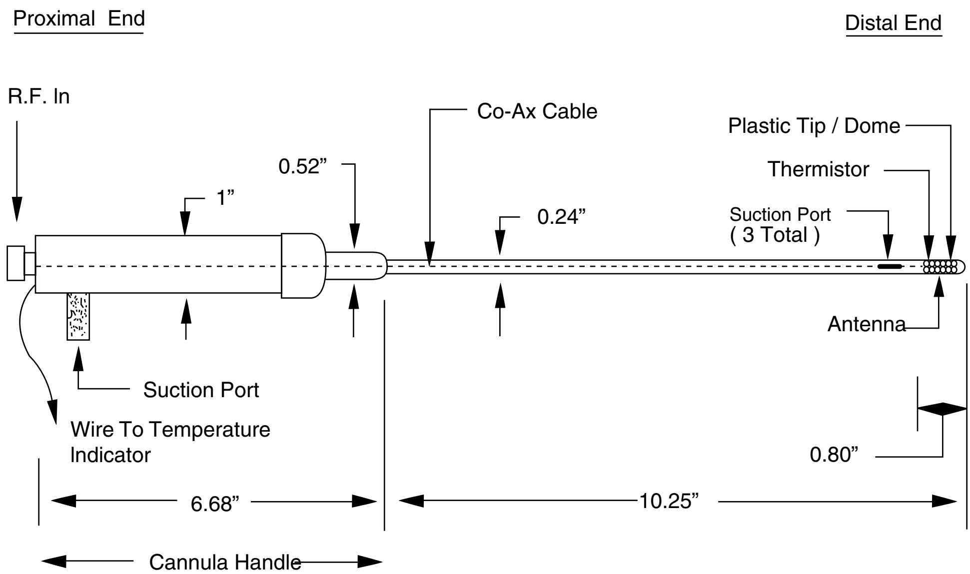


FIGURE 18.16 Microwave-aided liposuction cannula.

domain to compare the near field radiation patterns of several types of antennas including the dielectric tip monopole, open tip monopole, and metal tip monopole [67,68].

18.2.6 The Utilization of Biological Solder in Conjunction with Microwave Irradiation for Tissue Anastomoses in Endoscopic Surgery [69]

Endoscopic surgery is revolutionizing many surgical procedures. For example, laparoscopic surgical procedures, particularly laparoscopic cholecystectomy, have gained widespread acceptance. Further expansion of the endoscopic approach is inevitable. Although minimal access surgery is advantageous to patients, the technical problems imposed by the limited access are pushing existing tissue closure technologies (mechanical stapling devices and hand-sewn sutures) to their limits. The laparoscopic closure of an incision made in the bile duct for removal of stones is an example of the shortcomings of current technologies. Closure of this incision with laparoscopically placed sutures is difficult and post-operative bile leakage may result. Mechanical stapling devices for this purpose are beyond currently available technology.

To enhance a tissue anastomosis with microwaves, the tissue temperature must be kept below the threshold for damage, while the biological solder is heated above 60°C. Microwave anastomosis may also prove useful for vascular repairs, for example. A microwave antenna can be positioned inside an artery, and solder (albumin) is then placed on the outside of the vessel and in any small gaps between the arterial segments undergoing repair. The successful results *in vitro* have encouraged the preliminary investigation in a rabbit model. Early results *in vivo*, however, have indicated the need for a dry environment. More research is needed to evaluate the full potential of the microwave anastomosis technique.

18.2.7 Nerve Ablation for the Treatment of Gastroesophageal Reflux Disease

Gastroesophageal reflux disease (GERD) results from the chronic backward flow of stomach contents into the esophagus. The acid, bile, and digestive enzymes cause irritation of the esophagus and symptoms of heartburn, regurgitation, chest pain, voice disorders, and swallowing problems.

Normally, the muscular valve (lower esophageal sphincter or LES) at the junction of the esophagus and stomach prevents reflux from occurring. Reflux of stomach contents occurs when the LES and diaphragm are unable to provide enough tone or force to squeeze adequately on the esophagus. This may happen in some patients in whom the muscles have weakened over time or in those patients with hiatal hernias. The barrier function in these patients is completely lost, and reflux is present throughout the day.

The majority of patients with GERD, however, have normal LES and diaphragm pressures, yet the sphincter muscles relax frequently throughout the daytime to cause reflux. The relaxation events permit excessive reflux of stomach contents and the patient develops significant symptoms of GERD.

This abnormal event is a neurological reflex, termed transient lower esophageal relaxation (tLESR) [75–79], and is the cause of GERD in over 80% of patients. A tLESR is prompted when there is stretching of the stomach wall, as after a meal. The stretch receptors generate a nerve impulse, which travels upward within the myenteric plexus of the gastroesophageal junction. The myenteric plexus is a network of very small nerves lying between the layers of the stomach and esophagus musculature. The impulses travel through the LES, into the esophagus, and then join the vagus nerve on their way to the brain. When the brain receives these signals, a motor signal is sent to the LES causing prolonged relaxation.

There are hundreds of peer-reviewed scientific publications addressing the importance of tLESR in the development of GERD. Many investigators have collaborated to study the delivery of radio frequency energy for the treatment of GERD.

Investigators at Stanford [80] have recently performed radio frequency ablation of the stomach cardia (Figure 18.17a,b) in Yucatan mini-pigs to establish the effect on these nerve pathways. These nerve fibers course between the muscle layers of the LES and cardia. These investigators have demonstrated a statistically significant effect of delivering radio frequency energy to the cardia on the parameter of gastric yield pressure. This test is directly related to tLESRs. The stomach is stretched with carbon dioxide gas until the LES yields or relaxes in response to pressure. Yield pressures were higher in all animals after treatment, indicating that the nerve reflex arc was modulated to have a higher threshold for stimulation, or a lower frequency of transmission to the brain.

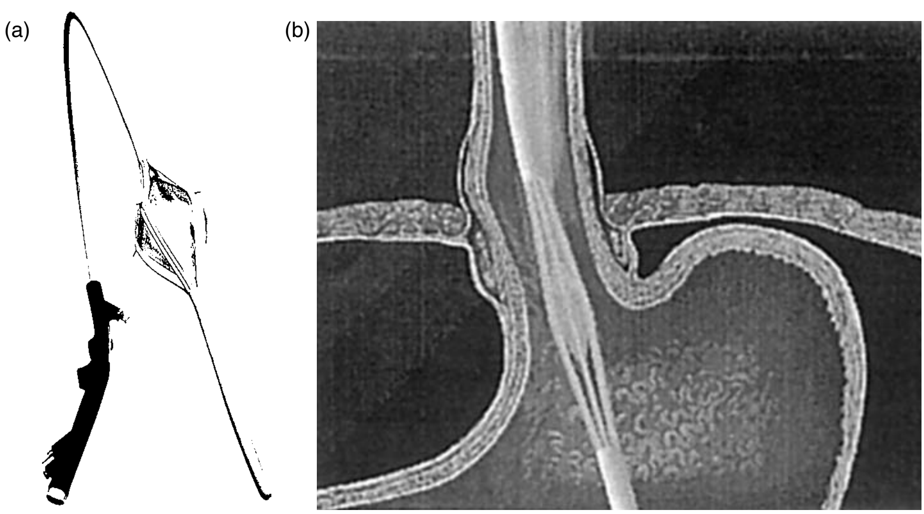


FIGURE 18.17 (a) Catheter used for nerve ablation in the treatment of gastroesophageal reflux disease; (b) Catheter positioning within the gastric cardia to deliver radiofrequency energy for nerve ablation.

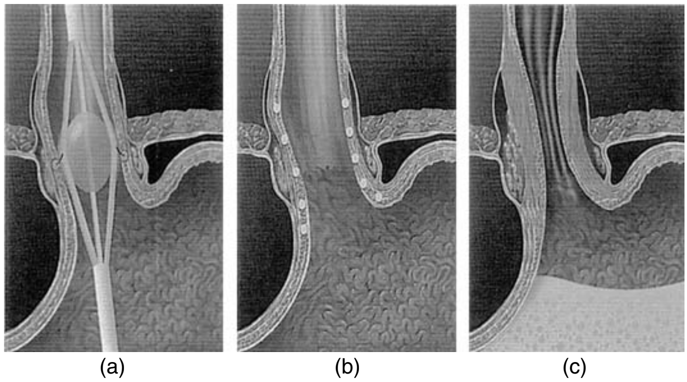


FIGURE 18.18 Step-by-step treatment for reflux.

18.2.7.1 Step-by-Step Treatment for Reflux—Research Carried Out at Conway Stuart Medical, Inc.

1. **Position, Inflate, Deploy, Irrigate, Treat.** The physician positions the catheter, inflates the balloon, deploys the needles and begins irrigation. During treatment, radio frequency energy is delivered in a controlled manner to the tissue surrounding the needle electrodes (Figure 18.18a).
2. **Treatment at multiple levels.** The treatment sequence is repeated to create well-defined coagulative lesions along the length of the lower esophageal sphincter and cardia (Figure 18.18b).
3. **Resorption and shrinking.** Over the next few weeks, the coagulated tissue resorbs and shrinks, increasing resistance to reflux (Figure 18.18c).

In an abstract entitled "Augmentation of lower esophageal sphincter pressure and gastric yield pressure after radiofrequency energy delivery to the lower esophageal sphincter muscle: a porcine model," Drs. D.S. Utley, M.A. Vierra, M.S. Kim, and G. Triadafilopoulos of VA Palo Alto Health Care System and Stanford University in Palo Alto, CA [80] report on their investigation of the technique of endoscopic, submucosal radio frequency energy (Rfe) delivery to the lower esophageal sphincter (LES) as a possible alternative treatment of GERD. Utilizing a porcine reflux model, they determined the effects of Rfe on LES pressure (LESP) and gastric yield pressure (GYP). In summary, Rfe is a promising new modality in the endoscopic treatment of GERD (Figure 18.18).

18.2.8 RF in the Treatment of Solid Organ Tumors

RITA Medical Systems, Inc. has developed a controlled tissue ablation system to treat solid organ tumors using minimally invasive RF ablation technology. This system includes an RF generator and a family of electrodes for the treatment of solid tumors. The controlled application of RF energy through an electrode placed directly in the tumor heats tissue to the required target temperature.

Each electrode consists of a thin hollow stainless steel shaft that acts as a primary electrode and also allows the introduction into the tumor of a curved array of secondary electrodes. The secondary electrodes have temperature sensors mounted on the tips to provide temperature feedback.

The RF generator delivers up to 50 watts to destroy or ablate the tumor. The operator sets the desired temperature; the generator automatically adjusts the power to attain the proper temperature and displays delivered power, impedance, and temperature.

The unique features of RF in treating solid tumors are as follows (Figure 18.19):

- *Minimally invasive* — many procedures can be performed through a laparascopic or even a percutaneous approach, frequently on an outpatient basis.

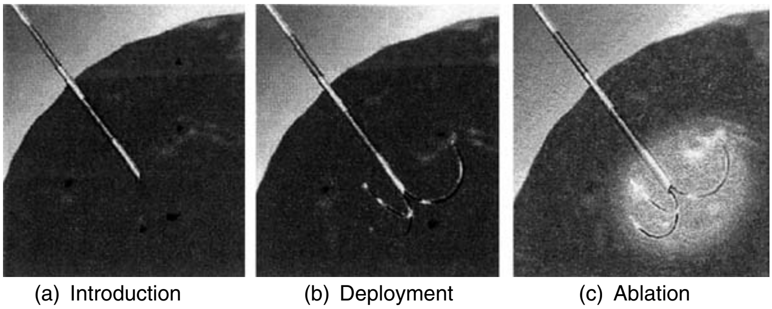


FIGURE 18.19 RF in the treatment of solid organ tumors. (With permission from RITA Medical Systems Inc.)

- *Creates large volume of ablated tissue* — the current device can ablate a spherical area of around 3 cm, approximating the size and shape of many cancerous lesions.
- *Temperature feedback leads to predictability and controllability*—the system provides temperature feedback at the periphery of the ablation volume to confirm tissue destruction. It also provides impedance feedback, which can be used to guide the application of RF power in order to ablate the tumor.

18.2.9 Application of RF Thermal Arthroscopy [72]

The combined use of RF energy and arthroscopy has been, in recent years, successfully advanced by ORATEC Interventions, Inc. They have accumulated an extensive collection of case reports, and from among these we will discuss one important clinical application. This method will serve as an example of the significance of RF thermal arthroscopy techniques.

In a report entitled, "Arthroscopic shoulder stabilization using suture anchors and capsular shrinkage," Jeffrey S. Abrams, M.D. of Princeton Orthopedic and Rehabilitation Associates, describes the treatment of recurrent shoulder joint instability after traumatic injury. The ligaments that normally anchor the upper arm to the shoulder joint can tear away from their attachments on the joint, (glenoid cavity) as a result of trauma. Without these ligamentous attachments (the labrum) the shoulder joint can repeatedly dislocate. Such recurrent instability not only eliminates recreational sports for the patients but can significantly affect their daily life [73].

The traditional approach to such injury involves orthopedic surgery to reattach the torn labrum and restore tension to the damaged ligaments, thus holding the joint together firmly. Laporoscopic techniques, in general, reduce surgical time, decrease morbidity, and can increase success rates. The arthroscopic technique for shoulder stabilization still involves the use of sutures to anchor the labrum back to the edge of the glenoid. RF electrothermal technology is then used to shrink the tissues of the ligamentous joint capsule and thus increase the tension on these ligaments. Increased tension on the ligaments, in turn, stabilizes the shoulder joint. Such procedures can be performed on an outpatient basis and require minimal postoperative medication.

18.3 Future Research

New therapeutic applications are continuously being developed that target organs heretofore untreated, utilizing the same general tools discussed in this chapter [81], and new companies are established to develop and market those technologies.

In this section of the chapter, the authors wish to address a number of topics for possible future research:

1. Microwave assisted photodynamic therapy (PDT) utilizing an LED/laser diffusing source contained inside a balloon onto which is imprinted a microwave antenna. The electromagnetic radiation from the antenna will provide tissue heating, thus increasing blood flow and tissue oxygenation needed for an effective photodynamic therapy [82].
2. Thermally molded biodegradable thermoplastic stent, containing a drug or drugs to retard restenosis, situated around a catheter-balloon. The catheter is to be placed in a vessel where microwave heating will facilitate the expansion of the stent against the vessel wall.
3. A catheter with distally located radiation generator, in which a microwave source, including an antenna or a light source, can be utilized for tissue ablation or local hyperthermia.
4. Dental instrument utilizing RF/microwave for treating teeth in which dental or endodontic instruments, such as drill or files, for example, are arranged to allow radiation of electromagnetic energy into a cavity in the tooth under treatment.

In addition, technology is presently being developed that allows for permanent implantation of RF/microwave wireless sensors in human subjects. We are currently researching and developing applications for permanently implantable pressure sensors and monitors.

18.4 Conclusions

In this chapter, we have reviewed a few of the existing applications of RF/microwaves in medicine. We have touched with some detail on the new applications currently under investigation. A more detailed discussion of some of the topics can be found in the book entitled *New Frontiers in Medical Device Technology* edited by Arye Rosen and Harel D. Rosen, published by John Wiley and Sons, 1995 as part of the Wiley Series in Microwave and Optical Engineering/Kai Chang, Series Editor.

Acknowledgments

We wish to recognize the assistance and advice of many who, since the early 1980s, have participated in research in the areas of RF/microwaves in medicine, some of whose research was covered in this chapter: from Jefferson Medical College, Drs. Paul Walinsky and Arnold J. Greenspon; from MMTC, Dr. Fred Sterzer and Mr. Dan Mawhinney; from Temple University, Dr. William Santamore; from the University of Pennsylvania, Dr. Louis Bucky. Gratitude is also due to Mr. John Hendrick of VidaMed, Inc., Mr. Hugh Sharkey of ORATEC, Interventions, Inc., and Mr. Barry Cheskin of RITA Medical Systems, Inc., who were so kind to furnish some of the material; and to Mr. Walter Janton for his technical skills and invaluable support. Finally, recognition is due to Mrs. Daniella Rosen for her contribution to the research on microwave-assisted liposuction, and for revising this manuscript again and again.

References

1. Cosman BJ, Cosman ER: *Guide to Radio Frequency Lesion Generation in Neurosurgery.* Radionics, Inc., 1974.
2. Cosman ER, Cosman BJ*: Methods of making nervous system lesions, Medical Therapy of Movement Disorders Guide to Radio Frequency Lesion Generation in Neurosurgery.* Radionics, Inc., 1974.
3. Alberts WW, Wright EW Jr, Feinstein B, Gleason CA: Sensory responses elicited by subcortical high frequency electrical stimulation in man. *J. Neurosurg.,* 36, 80–82, 1972.
4. Dieckmann G, Gabriel E, Hassler R: Size, form, and structural peculiarities of experimental brain lesions obtained by thermocontrolled radiofrequency. *Confin. Neurol.,* 26, 134–142, 1965.
5. Brodkey JS, Miyazaki Y, Ervin FR, Mark VH: Reversible heat lesions with radiofrequency current: A method of stereotactic localization. *J. Neurosurg.,* 21, 49–53, 1964.
6. Fager CA: Surgical treatment of involuntary movement disorders, *Lahey Clin. Found. Bull.,* 22, 79–83, 1973.

7. Hurt RW, Ballantine HT Jr: Stereotactic anterior cingulate lesions for persistent pain: A report on 68 cases. *Clin. Neurosurg.,* 21, 334–351, 1974.
8. Schwan HP: Electrical properties of tissues and cell suspensions, *Advanced Phys. Med. Biol.,* 5, 147–209, 1957.
9. Stuchly MA, Stuchly SS: Dielectric properties of biological substances-tabulated, *J. Microwave Power,* 15, 19–26, 1980.
10. Foster KR, Schepps JL, Schwan HP: Microwave dielectric relaxation in tissue: a second look, *Biophys. J.,* 29, 271–281, 1980.
11. Thuari M, Steel MC, Sheppard RJ, Grant HE: Dielectric properties of developing rabbit brain at 37 degrees C, *Bioelectromagnetics,* 6, 235–242, 1985.
12. Schwan HP, Foster KR: RF-field interactions with biological systems: electrical properties and biophysical mechanism, *Proc. IEEE,* 68, 104–113, 1980.
13. Pethig R, Kell DB: The passive electrical properties of biological systems: their significance in physiology, biophysics, and biotechnology, *Phys. Med. Biol.,* 32, 933–970, 1987.
14. Duck, FA, *Physical Properties of Tissue: A Comprehensive Reference Book,* Academic Press, New York, 1990.
15. Gabriel C, Gabriel S, Corthout E: The dielectric properties of biological tissues: I. Literature review, *Phys. Med. Biol.,* 41, 2231–2249, 1996.
16. Grant HE, Sheppard RJ, South FP: *Dielectric Behaviour of Biological Molecules in Solution,* Oxford University Press, Oxford, 1978.
17. Foster KR, Schepps JL: Dielectric properties of tumor and normal tissues at radio through microwave frequencies, *J. Microwave Power,* 16, 107–119, 1981.
18. Gabriel S, Lau RW, Gabriel C: The dielectric properties of biological tissues: II. Measurements in the frequency range 10 Hz to 20 Ghz, *Phys. Med. Biol.,* 41, 2251–2269, 1996.
19. Cole KS, Cole RH: Dispersion and absorption in dielectrics: I. Alternating current characteristics, *J. Chem. Phys.,* 9, 341–351, 1941.
20. Debye P, Huckel E: *Phys. Z.,* 24, 185, 1923.
21. Frohlich H: *Theory of Dielectrics*, Oxford University Press, Oxford, England, 1949.
22. Daniel VV: *Dielectric Relaxation,* Academic Press, London, 1967.
23. Sterzer F et al.: RF therapy for malignancy, *IEEE Spectrum,* 17, 12, 32–37, Dec. 1980.
24. Lele, PP: Induction of deep, local hyperthermia by ultrasound and electromagnetic fields: problems and choices. *Radiat. Environ. Biophy.,* 17, 205, 1980.
25. Tofighi M-R: A two-port microstrip test fixture for measurement of complex permittivity of biological tissues at microwave and millimeter wave frequencies, PhD. Dissertation, Department of ECE, Drexel University, Philadelphia, PA.
26. Roberts S, Von Hippel A: A new method for measuring dielectric constant and loss in the range of centimeter waves, *J. Appl. Phys.,* 17, 610–616, 1946.
27. Westphal: Dielectric measuring techniques, in *Dielectric Materials and Applications,* A.R. von Hippel, Ed., Wiley, New York, 63–122, 1954.
28. Burdette EC, Cain FL, Seals J: *In vivo* probe measurement technique for determining dielectric properties at VHF through microwave frequencies, *IEEE Trans. Microwave Theory and Tech.,* MTT-28, 414–424, 1980.
29. Stuchly SS, Sibbald CL, Anderson JM: A new aperture admittance model for open-ended waveguides, *IEEE Trans. Microwave Theory and Tech.,* MTT-42, 192–198, 1994.
30. Steel MC, Sheppard RJ, The dielectric properties of rabbit tissue, pure water and various liquids suitable for tissue phantoms at 35 Ghz, *Phys. Med. Biol.,* 33, 467–472, 1988.
31. Steel MC, Sheppard RJ, Collin R: Precision waveguide cells for the measurement of complex permittivity of lossy liquids and biological tissue at 35 Ghz, *Journal of Physics. E. Scientific Instruments,* 20, 872–877, 1987.
32. Land DV, Campbell AM: A quick accurate method for measuring the microwave dielectric properties of small tissue samples, *Phys. Med. Biol.,* 37, 183–192, 1992.

33. Athey TW, Stuchly MA, Stuchly SS: Measurment of radio frequency permittivity of biological tissues with an open-ended coaxial line: part I, *IEEE Trans. Microwave Theory and Tech.,* MTT-30, 82–86, 1982.
34. Kraszewski A, Stuchly MA, Stuchly SS, Smith M: *In vivo* and *in vitro* dielectric properties of animal tissues at radio frequencies, *Bioelectromagnetics,* 3, 421–432, 1982.
35. Nyshadham A, Sibbald CL, Stuchly SS, Permittivity measurements using open-ended sensors and reference liquid calibration—an uncertainty analysis, *IEEE Trans. Microwave Theory and Tech.,* MTT-40, 305–314, 1992.
36. Misra DM, Chabbra M, Epstein BR, Mirotznik M, Foster KR: Noninvasive electrical characteriation of materials at microwave frequencies using an open-ended coaxial line: test of an improved calibration technique, *IEEE Trans. Microwave Theory and Tech.,* MTT-38, 8–13, 1990.
37. Gabriel C, Chan TYA, Grant EH: Admittance models for open ended coaxial probes and their place in dielectric spectroscopy, *Phys. Med. Biol.,* 39, 2183–2199, 1994.
38. Bao J, Lu S, Hurt DW: Complex dielectric measurements and analysis of brain tissues in the radio and microwave frequencies, *IEEE Trans. Microwave Theory and Tech.,* MTT-45, 1730–1740, 1997.
39. Daryoush A, private communications.
40. King RWP, Smith GS: *Antennas in Matter,* The MIT Press, Cambridge, MA, 1981.
41. King RWP, Trembly BS, Strohbehn JW: The electromagnetic field of an insulated antenna in a conducting or dielectric medium, *IEEE Trans. on Microwave Theory and Tech.*, MTT-31, 7, 574–583, July 1983.
42. Iskander MF, Umeh AM: Design optimization of interstitial antennas, *IEEE Trans. Biomed. Eng.*, 36, 238–246, Feb. 1989.
43. Rumeh AM, Iskander MF: Performance comparison of available interstitial antennas for microwave hyperthermia, *IEEE Trans. on Microwave Theory and Tech.*, 37, 7, 1126–1133, July 1989.
44. Debicki PS, Astrahan MA: Calculating input impedance of electrically small insulated antennas for microwave hyperthermia, *IEEE Trans. on Microwave Theory and Techniques,* 41, 2, 357–360, February 1993.
45. Su DW-F, Wu L-K, Input impedance characteristics of coaxial slot antennas for interstitial microwave hyperthermia, *IEEE Trans. on Microwave Theory and Tech.,* 47, 3, 302–307, March 1999.
46. Casey JP, Bansal R: The near field of an insulated dipole in a dissipative dielectric medium, *IEEE Trans. on Microwave Theory and Tech.*, 34, 4, 459–463, April 1986.
47. Greenspon AJ, Walinsky P, Rosen A: Catheter ablation for the treatment of cardiac arrhythmias, in *New Frontiers in Medical Technology,* John Wiley & Sons, Inc., New York, 1995.
48. Rosenbaum RM, Greenspon AJ, Hsu S, Walinsky P, Rosen A: RF and microwave ablation for the treatment of ventricular tachycardia, *IEEE MTT-S Digest,* 1993.
49. Sterzer F: Localized hyperthermia treatment of cancer, *RCA Rev.,* 42, 727, 1981.
50. Hahn GM: *Hyperthermia and Cancer,* Plenum Press, New York, 1982.
51. Storm FK: *Hyperthermia in Cancer Therapy.* GK Hall, Boston, 1983.
52. Walinsky P, Rosen A, Greenspon AJ: Method and apparatus for high frequency catheter ablation. U.S. Pat. 4,641,649.
53. Walinsky P, Rosen A, Martinez-Hernandez A, Smith D, Nardone DO, Brevette B: Microwave balloon angioplasty, *The J. of Invasive Cardiology,* 3, 3, May/June 1991.
54. Langberg JJ, Wonnell TL, Chin M, et al.: Catheter ablation of the atrioventricular junction using a helical microwave antenna: A novel means of coupling energy to the endocardium. *PACE,* 14, 2105, 1991.
55. Wonnell TL, Stauffer PR, Langberg JJ: Evaluation of microwave and RF catheter ablation in a myocardial equivalent phantom model. *IEEE Trans. Biomed. Eng.*, 39, 1086, 1992.
56. Rosen A, Walinsky P, Smith D, Kosman Z, Martinez A, Sterzer F, Presser A, Mawhinney D, Chou J-S, Goth P: Studies of microwave thermal balloon angioplasty in rabbits, *IEEE MTT-S Digest,* 1993.
57. Rosen A, Rosen HD: The efficacy of transurethral thermal ablation in the management of benign prostatic hyperplasia, in *New Frontiers in Medical Technology,* John Wiley & Sons, Inc., New York, NY, 1995.

58. Rosen A, Walinsky P: Microwave balloon angioplasty, in *New Frontiers in Medical Technology,* John Wiley & Sons, Inc., New York, NY, 1995.
59. Rosen HD, Rosen A: RF/Microwaves, a hot topic in medicine, *IEEE Potentials,* August/September 1999.
60. Sterzer F: Localized heating of deep-seated tissues using microwave balloon catheters, in *New Frontiers in Medical Technology,* John Wiley & Sons, Inc., New York, 1995.
61. Schmidt-Nowara W et al.: Oral appliances for the treatment of snoring and obstructive sleep apnea, *Sleep,* 18, 6, 501–510, 1995.
62. Simmons FB, Guilleminault C, Miles LE: The platopharyngoplasty operation for snoring and sleep apnea: An interim report. *Otolaryngol. Head Neck Surg.,* 92, 375–380, 1984 and Conway W, Fujita S, Zorick F, et al.: Uvulopalatopharyngoplasty: One year follow-up. *Chest,* 88, 385–387, 1985.
63. Riley et al. 1995, op. cit. The description of this approach and the outcomes data are taken from this paper, though the team has published several papers on the method since 1988.
64. Powell NB, Riley RW, Troell RJ, Blumen MB, and Guilleminault C: Radiofrequency volumetric reduction of the tongue — A porcine pilot study for the treatment of obstructive sleep apnea syndrome, laboratory and animal investigations, *Chest,* 111, 1348–55, 1997.
65. Rosen A, Rosen D, Tuma G, Bucky L,: RF/Microwave aided tumescent liposuction, *IEEE MTT Transactions,* (Nov. 2000), in press.
66. Jemison W et al.: New antenna design for microwave assisted liposuction, to be published; Private communications.
67. Hurter W, Reinbold F, Lorenz WJ: A dipole antenna for interstitial microwave hyperthermia, *IEEE Trans. on Microwave Theory and Techniques,* 34, 4, 459–463, April 1986.
68. Labonte S, Blais A, Legault SR, Ali HO, Roy L: Monopole antennas for microwave catheter ablation, *IEEE Trans. on Microwave Theory and Techniques,* 44, 10, 1832–1840, Oct. 1996.
69. Santamore W: Private communication.
70. *Health aspects of radio frequency and microwave radiation exposure, Part 1*, 77-EHD-13, National Health and Welfare, Canada, November 1977.
71. Paglione R: Medical applications of microwave energy, *RCA Engineer,* 27, 5, 17–21, Sept./Oct. 1982.
72. Sharkey H: Private communication.
73. Abrams JS: Arthroscopic shoulder stabilization using suture anchors and capsular shrinkage, ORATEC Interventions, Inc., Applications in Electrothermal Arthroscopy, Case Report.
74. Tell RA: Microwave Energy Absorption in Tissue, *Techn. Rep. PB*, Environmental Protection Agency, Feb. 1972.
75. Mittal RK, Holloway RH, Penagini R, Blackshaw LA, Dent D: Transient lower esophageal sphincter relaxation. *Gastroenterology* 1995; 109:601–610.
76. Rawahara H, Dent J, Davidson G: Mechanisms responsible for gastroesophageal reflux in children. *Gastroenterology* 1997; 113:399–408.
77. Panagini R, Bianchi PA: Effect of morphine on gastroesophageal reflux and transient lower esophageal sphincter relaxation. *Gastroenterology* 1997; 113:409–414.
78. Blackshaw LA, Haupt JA, Omari T, Dent J: Vagal and sympathetic influences on the ferret lower esophageal sphincter. *J Autonomic Nerv System* 1997; 66:179–188.
79. Stakeberg J, Lehman A: Influence of different intragastric stimuli on triggering of transient lower esophageal sphincter relaxation in the dog. *Neurogastroenterol* 1999; 11:125–132.
80. Private communications with Edwards SD.
81. Andre Vander Vorst, Arye Rosen, Youji Kotsuka: *RF/Microwave Interaction with Biological Tissues,* Wiley Press, January 2006.
82. Rosen A, Rosen HD, Hsi RA, Rosen D, Happawana G, Evans G, Fathy A, Stern L: Advances in RF/Microwave and Light Source for Applications in Therapeutic Medicine, Conference Proceedings, Volume 2, XV International Conference on Microwaves, Radar and Wireless Communciations, Poland, Warszawa, May 17–19 2004.

II

Systems Considerations

19

Thermal Analysis and Design of Electronic Systems

Avram Bar-Cohen
University of Maryland

Karl J. L. Geisler
General Dynamics—Advanced Information Systems

Allan D. Kraus
Allan D. Kraus Associates

19.1 Motivation

In the thermal control of radio frequency (RF) devices, it is necessary to provide an acceptable microclimate for a diversity of devices and packages that vary widely in size, power dissipation, and sensitivity to temperature. Although the thermal management of all electronic components is motivated by a common set of concerns, this diversity often leads to the design and development of distinct thermal control systems for different types of electronic equipment. Moreover, owing to substantial variations in the performance, cost, and environmental specifications across product categories, the thermal control of similar components may require widely differing thermal management strategies.

The prevention of catastrophic thermal failure (defined as an immediate, thermally induced, total loss of electronic function) must be viewed as the primary and foremost aim of electronics thermal control. Catastrophic failure may result from a significant deterioration in the performance of the component/system or from a loss of structural integrity at the relevant packaging levels. In early microelectronic systems, catastrophic failure was primarily *functional* and thought to result from changes in the bias voltage, *thermal runaway* produced by regenerative heating, and dopant migration, all occurring at elevated transistor junction temperatures. Although these failure modes may still occur during the device development process, improved semiconductor simulation tools and thermally compensated devices have largely quieted these concerns and substantially broadened the operating temperature range of today's RF devices.

In microelectronic, microwave, and RF components, the levels of integration and device density on the chips, as well as frequencies of operation, continue to increase. The most critical heat-producing component for most RF systems is the power amplifier (PA) stage. Output power required from these stages ranges from less than 1 W for some handheld commercial applications to greater than 1 kW (multiple parallel stages) for certain military, avionics, and data link applications. Single transistor output power levels are as high as 100–200 W for applications ranging from commercial base stations to avionics, satellite communications, and military. Amplified efficiencies for the highest output power requirements are typically in the range of 15–35%. To facilitate effective thermal management for such high power levels, PA operation must be pulsed with low duty cycles (reducing thermal power dissipation requirements). Improved thermal performance can be translated into higher duty cycles and therefore into greater data transfer or more efficient use of bandwidth. For these kinds of applications, performance is already limited primarily by the maximum achievable heat flux. Improvements in that figure-of-merit automatically and immediately translate into improved system performance.

More generally, however, thermal design is aimed at preventing thermally induced physical failures through reduction of the temperature rise above ambient and minimization of temperature variations within the packaging structure(s). With RF-integrated circuits or discrete RF high-performance devices, maximum frequency of operation, noise figure, power saturation levels, and nonlinear behavior are all affected by temperature. The use of many low-temperature materials and the structural complexity of chip packages and printed circuit boards (PCB) has increased the risk of catastrophic failures associated with the vaporization of organic materials, the melting of solders, and thermal-stress fractures of leads, joints, and seals as well as the fatigue-induced delamination and fracture or creep-induced deformation of encapsulants and laminates. To prevent catastrophic thermal failure, the designer must know the maximum allowable temperatures, acceptable internal temperature differences, and the power consumption/dissipation of the various components. This information can be used to select the appropriate fluid, heat-transfer mode, and inlet temperature for the coolant and to thus establish the thermal control strategy early in the design process.

After the selection of an appropriate thermal control strategy, attention can be turned to meeting the desired system-level reliability and the target failure rates of each component and subassembly. Individual solid-state electronic devices are inherently reliable and can typically be expected to operate, at room temperature, for some 100,000 years, that is, with a base failure rate of 1 FIT (failures in 10^9 h). However, since the number of devices in a typical radio component is rapidly increasing and since an RF system may consist of many tens to several hundreds of such components, achieving a system Mean Time Between Failures of several thousand hours in military equipment and 40,000–60,000 h in commercial systems is a most formidable task.

Many of the failure mechanisms, which are activated by prolonged operation of electronic components, are related to the local temperature and/or temperature gradients, as well as the thermal history of the package [1]. Device-related functional failures often exhibit a strong relationship between failure rate and operating temperature. This dependence can be represented in the form of an exponential Arrhenius relation, with unique, empirically determined coefficients and activation energy for each component type and failure mechanism. In the normal operating range of microelectronic components, a 10–20°C increase in chip temperature may double the component failure rate, and even a 1°C decrease may then lower the predicted failure rate associated with such mechanisms by 2–4% [2].

Unfortunately, it is not generally possible to characterize thermally induced structural failures, which develop as a result of differential thermal expansion among the materials constituting a microwave package, in the form of an Arrhenius relation. Although these mechanical stresses may well increase as the temperature of the component is elevated, thermal stress failures are, by their nature, dependent on the details of the local temperature fields, as well as the assembly, attachment, and local operating history of the component. Furthermore, thermal-stress generation in packaging materials and structures is exacerbated by power transients, as well as by the periodically varying environmental temperatures, experienced by most electronic systems, during both qualification tests and actual operation. However, stress variations in the elastic domain or in the range below the fatigue limit may have little effect on the component failure rate. Consequently, the minimization of elimination of thermally induced failures often requires careful attention to both the temperature and stress fields in the electronic components and necessitates the empirical validation of any proposed thermostructural design criteria.

19.1.1 Thermal Packaging Options

When the heat flux dissipated by the electronic component, device, or assembly is known and the allowable temperature rise above the local ambient condition is specified, the equations of the following sections can be used to determine which heat transfer process or combination of processes (if any) can be employed to meet the desired performance goals. Figure 19.1 shows the variation of attainable temperature differences with surface heat flux for a variety of heat transfer modes and coolant fluids. Examination of Figure 19.1 reveals that for a typical allowable temperature difference of 60°C between the component surface and the ambient, "natural" cooling in air—relying on both free convection and radiation—is effective only for heat fluxes below approximately 0.05 W/cm^2. Although forced convection cooling in air offers approximately an order-of-magnitude improvement in heat transfer coefficient, this thermal configuration is unlikely to provide heat removal capability in excess of 1 W/cm^2 even at an allowable temperature difference of 100°C.

To facilitate the transfer of moderate and high heat fluxes from component surfaces, the thermal designer must choose between the use of finned, air-cooled heat sinks, and direct or indirect liquid cooling. Finned arrays and sophisticated techniques for improving convective heat transfer coefficients can extend the effectiveness of air cooling to progressively higher component heat fluxes, but often at ever-increasing weight, cost, and volume penalties. Alternately, reliance on heat transfer to liquids flowing at high velocity through so-called cold plates can offer a dramatic improvement in the transferable

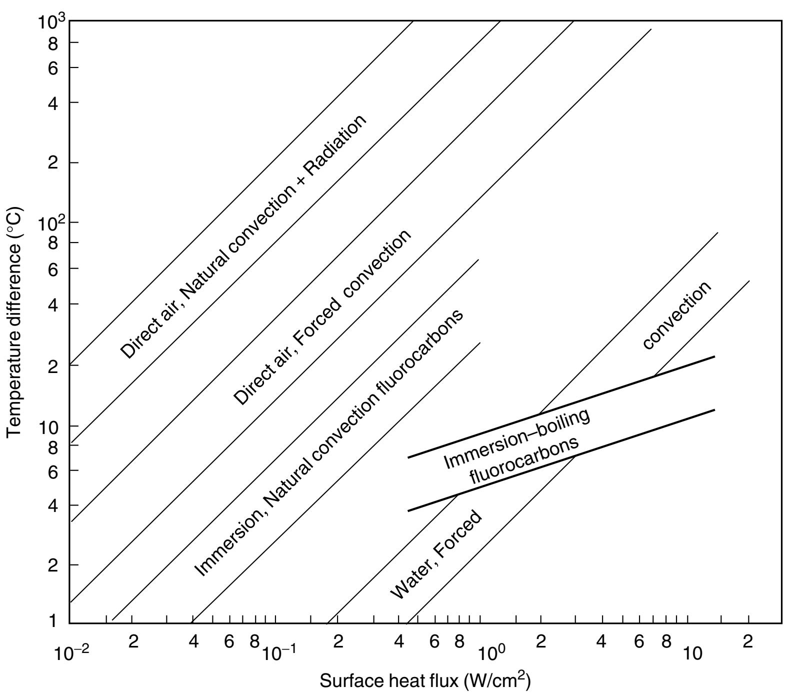


FIGURE 19.1 Temperature differences attainable as a function of heat flux for various heat transfer modes and various coolant fluids. (From A.D. Kraus and A. Bar-Cohen, *Thermal Analysis and Control of Electronic Equipment*, New York: McGraw-Hill, 1983. With permission.)

heat flux even at temperature differences as low as 10°C, when the conduction resistance in the cold plate wall is negligible.

A similar high heat flux capability is offered by boiling heat transfer to perfluorinated (FCs) and hydrofluoroether (HFEs) liquids. The high dielectric strength and low dielectric constant of these liquids make it possible to implement this approach for a wide range of components. Direct liquid contact allows the removal of component heat fluxes in excess of 10 W/cm^2 with saturated pool boiling at temperature differences typically less than 20°C. Natural convection (i.e., nonboiling) immersion cooling can also offer significant advantages and, as seen in Figure 19.1, serves to bridge the gap between direct air cooling and cold plate technology.

Unfortunately, when addressed within stringent cost targets, the cooling requirements of twenty-first century microelectronic, microwave, and RF components cannot be met by today's thermal packaging technology. Rather, ways must be sought to improve on currently available technology, to leverage and combine the best features of existing thermal packaging hardware, and to introduce unconventional, perhaps even radical, thermal solutions into the electronic product family. In so doing, attention must be devoted to three primary issues:

- Highly effective air cooling—removing dissipated power from one or several high-performance components within minimal volumes and with low air-side pressure drops.
- Heat spreading—transporting heat from the relatively small area of the device to the substrate, card, or board, or to a relatively large heat sink or cold plate base.
- Interfacial heat transfer—transferring heat across the thermal resistances between the device and the next level of thermal packaging.

Attention now turns to a detailed discussion of basic heat transfer and the determination of the various types of thermal resistances often encountered in electronic equipment.

19.2 Thermal Modeling

To determine the temperature differences encountered in the flow of heat within electronic systems, it is necessary to recognize the relevant heat transfer mechanisms and their governing relations. In a typical system, heat removal from the active regions of the device(s) may require the use of several mechanisms, some operating in series and others in parallel, to transport the generated heat to the coolant or ultimate heat sink. Practitioners of the thermal arts and sciences generally deal with four basic thermal transport modes: conduction, convection, phase change, and radiation.

19.2.1 Conduction Heat Transfer

19.2.1.1 One-Dimensional Conduction

Steady thermal transport through solids is governed by the Fourier equation, which in one-dimensional form, is expressible as

$$q = -kA\frac{\mathrm{d}T}{\mathrm{d}x} \tag{19.1}$$

where q is the heat flow, k is the thermal conductivity of the medium, A is the cross-sectional area for the heat flow, and $\mathrm{d}T/\mathrm{d}x$ is the temperature gradient. As depicted in Figure 19.2, heat flow produced by a negative temperature gradient is considered positive. This convention requires the insertion of the minus sign in Equation 19.1 to assure a positive heat flow, q. The temperature difference resulting from

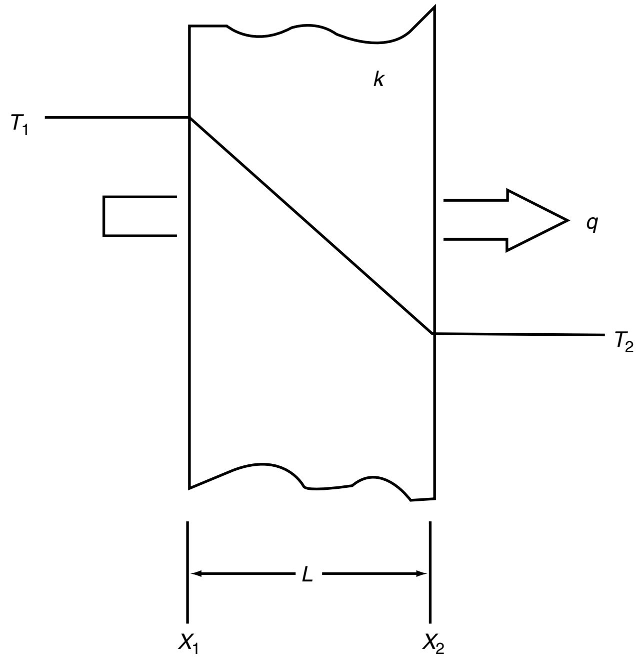


FIGURE 19.2 One-dimensional conduction through a slab. (From A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995. With permission.)

the steady-state diffusion of heat is thus related to the thermal conductivity of the material, the cross-sectional area, and the path length, L, according to

$$(T_1 - T_2)_{\text{cd}} = q\frac{L}{kA} \tag{19.2}$$

The form of Equation 19.2 suggests that, by analogy to Ohm's Law governing electrical current flow through a resistance, it is possible to define a thermal resistance for conduction, R_{cd}, as

$$R_{\text{cd}} \equiv \frac{(T_1 - T_2)}{q} = \frac{L}{kA} \tag{19.3}$$

19.2.1.2 One-Dimensional Conduction with Internal Heat Generation

Situations in which a solid experiences internal heat generation, such as that produced by the flow of an electric current, give rise to more complex governing equations and require greater care in obtaining the appropriate temperature differences. The axial temperature variation in a slim, internally heated conductor whose edges (ends) are held at a temperature T_{o} is found to equal

$$T = T_{\text{o}} + q_{\text{g}}\frac{L^2}{2k}\left[\left(\frac{x}{L}\right) - \left(\frac{x}{L}\right)^2\right] \tag{19.4}$$

When the volumetic heat generation rate, q_{g} in W/m^3, is uniform throughout, the peak temperature is developed at the center of the solid and is given by

$$T_{\max} = T_{\text{o}} + q_{\text{g}}\frac{L^2}{8k} \tag{19.5}$$

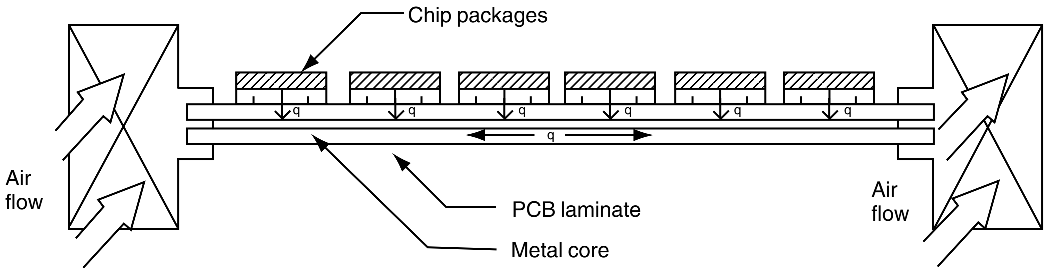


FIGURE 19.3 Edge-cooled printed circuit board populated with components. (From A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995. With permission.)

Alternatively, since q_g is the volumetric heat generation, $q_g = q/LW\delta$, the center-edge temperature difference can be expressed as

$$T_{\max} - T_o = q\frac{L^2}{8kLW\delta} = q\frac{L}{8kA} \tag{19.6}$$

where the cross-sectional area, A, is the product of the width, W, and the thickness, δ. An examination of Equation 19.6 reveals that the temperature difference in a conductor with a distributed heat input is only one-half that of a structure in which all of the heat is generated at the center.

In the design of airborne electronic systems and equipment to be operated in a corrosive or damaging environment, it is often necessary to conduct the heat dissipated by the components down into the substrate or PCB and, as shown in Figure 19.3, across the substrate/PCB to a cold plate or sealed heat exchanger. For a symmetrically cooled substrate/PCB with approximately uniform heat dissipation on the surface, a first estimate of the peak temperature at the center of the board, can be obtained using Equation 19.6.

This relation can be used effectively in the determination of the temperatures experienced by conductively cooled substrates and conventional PCBs, as well as PCBs with copper lattice on the surface, metal cores, or heat sink plates in the center. In each case it is necessary to evaluate or obtain the effective thermal conductivity of the conducting layer. As an example, consider an alumina substrate, 0.20 m long, 0.15 m wide, and 0.005 m thick with a thermal conductivity of 20 W/mK, whose edges are cooled to 35°C by a cold plate. Assuming that the substrate is populated by 15 compounds, each dissipating 2 W, the substrate center temperature will be found to be equal to 85°C when calculated using Equation 19.6.

19.2.1.3 Spreading Resistance

In component packages that provide for lateral spreading of the heat generated in the device(s), the increasing cross-sectional area for heat flow at successive "layers" adjacent to the device reduces the internal thermal resistance. Unfortunately, however, there is an additional resistance associated with this lateral flow of heat. This, of course, must be taken into account in the determination of the overall component package temperature difference.

For the circular and square geometries common in many applications, Negus et al. [3] provided an engineering approximation for the spreading resistance of a small heat source on a thick substrate or heat spreader (required to be 3–5 times thicker than the square root of the heat source area), which can be expressed as

$$R_{\mathrm{sp}} = \frac{0.475 - 0.62\epsilon + 0.13\epsilon^2}{k\sqrt{A_c}} \tag{19.7}$$

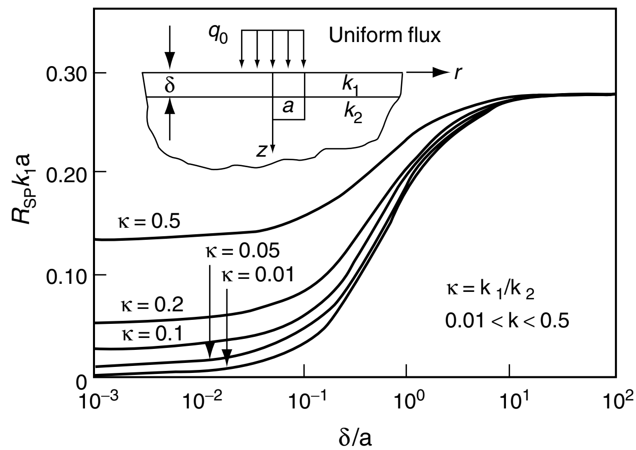


FIGURE 19.4 The thermal resistance for a circular heat source on a two-layer substrate. (From M.M. Yovanovich and V.W. Antonetti, In *Advances in Thermal Modeling of Electronic Components and Systems*, A. Bar-Cohen and A.D. Kraus (eds.), New York: Hemisphere Publishing Co., 79–128, 1988. With permission.)

where ϵ is the ratio of the heat source area to the substrate area, k is the thermal conductivity of the substrate, and A_c is the area of the heat source.

For relatively thin layers on thicker substrates, such as encountered in the use of thin lead frames, or heat spreaders interposed between the device and substrate, Equation 19.7 cannot provide an acceptable prediction of R_{sp}. Instead, use can be made of the numerical results plotted in Figure 19.4 to obtain the requisite value of the spreading resistance.

19.2.1.4 Interface/Contact Resistance

Heat transfer across the interface between two solids is generally accompanied by a measurable temperature difference, which can be ascribed to a contact or interface thermal resistance. For perfectly adhering solids, geometrical differences in the crystal structure (lattice mismatch) can impede the flow of photons and electrons across the interface, but this resistance is generally negligible in engineering design. When dealing with real interfaces, the asperities present on each of the surfaces, as shown in an artist's conception in Figure 19.5, limit actual contact between the two solids to a very small fraction of the apparent interface area. The flow of heat across the gap between two solids in nominal contact is, thus, seen to involve solid conduction in the areas of actual contact and fluid conduction across the "open" spaces. Radiation across the gap can be important in a vacuum environment or when the surface temperatures are high.

The total contact conductance, h_{co} is taken as the sum of the solid-to-solid conductance, h_c and the gap conductance, h_g

$$h_{co} = h_c + h_g \tag{19.8}$$

The contact resistance based on the apparent contact area, A_a may be defined as

$$R_{co} \equiv \frac{1}{h_{co}A_a} \tag{19.9}$$

In Equation 19.8, h_c is given by Yovanovich and Antonetti [4] as

$$h_c = 1.25k_s\left(\frac{m}{\sigma}\right)\left(\frac{P}{H}\right)^{0.95} \tag{19.10}$$

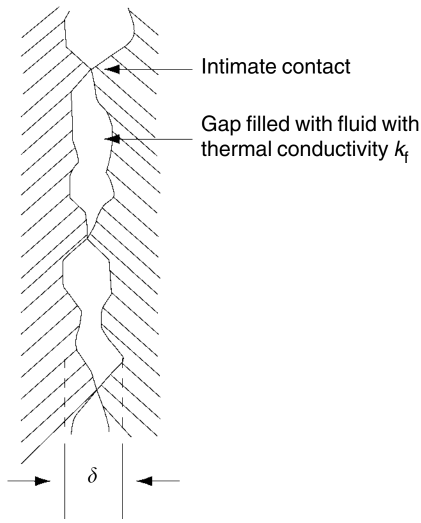


FIGURE 19.5 Physical contact between two nonideal surfaces. (From A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995. With permission.)

where P is the contact pressure and H is the micro-hardness of the softer material (both in Pa), k_s is the harmonic mean thermal conductivity for the two solids with thermal conductivities, k_1 and k_2,

$$k_s = \frac{2k_1 k_2}{k_1 + k_2}$$

σ is the effective rms surface roughness developed from the surface roughnesses of the two materials, σ_1 and σ_2,

$$\sigma = \sqrt{\sigma_1^2 + \sigma_2^2}$$

and m is the effective absolute surface slope composed of the individual slopes of the two materials, m_1 and m_2,

$$m = \sqrt{m_1^2 + m_2^2}$$

In the absence of detailed information, the σ/m ratio can be assumed to fall in the range of 5–9 μm for relatively smooth surfaces [5].

For normal interstitial gases around atmospheric pressure, h_g in Equation 19.8 is given by

$$h_g = \frac{k_g}{Y} \tag{19.11}$$

where k_g is the thermal conductivity of the gap fluid, and Y is the distance between the mean planes given by

$$Y = 1.185\left[-\ln\left(3.132\frac{P}{H}\right)\right]^{0.547}\sigma$$

Equations 19.10 and 19.11 can be added, and then in accordance with Equation 19.9, the total contact resistance becomes

$$R_{\text{co}} \equiv \left\{ \left[1.25 k_{\text{s}} \left(\frac{m}{\sigma}\right) \left(\frac{P}{H}\right)^{0.95} + \frac{k_{\text{g}}}{Y} \right] A_{\text{a}} \right\}^{-1} \tag{19.12}$$

19.2.1.5 Transient Heating or Cooling

An externally heated solid of relatively high thermal conductivity that is experiencing no external cooling, will undergo a constant rise in temperature according to

$$\frac{\mathrm{d}T}{\mathrm{d}x} = \frac{q}{mc_{\text{p}}} \tag{19.13}$$

where q is the rate of internal heat generation, m is the mass of the solid, and c_{p} is the specific heat of the solid. Equation 19.13 assumes that all of the mass can be represented by a single temperature and this relation is frequently termed the “lumped capacity” solution for transient heating.

Expanding on the analogy between thermal and electrical resistances, the product of mass and specific heat can be viewed as analagous to electrical capacitance and thus to consitute the “thermal capacitance.”

When the same solid is externally cooled, the temperature rises asymptotically toward the steady-state temperature, which is itself determined by the external resistance to the heat flow, R_{ex} Consequently, the time variation of the temperature of the solid is expressible as

$$T(t) = T(t-0) + qR_{\text{ex}}[1 - \mathrm{e}^{-t/mc_{\text{p}}R_{\text{ex}}}] \tag{19.14}$$

The lumped capacitance model is accurate when the ratio of the internal conduction resistance of a solid to the external thermal resistance is small. This ratio is represented by the Biot number, and the criterion for applicability of the lumped capacitance model is given as

$$\text{Bi} = \frac{hL_{\text{c}}}{k} < 0.1 \tag{19.15}$$

where the characteristic length, L_{c}, is typically defined as the ratio of the solid's volume to its surface area. More generally, L_{c} should be taken as the distance over which the solid experiences its maximum temperature difference [6].

19.2.2 Convective Heat Transfer

19.2.2.1 The Heat Transfer Coefficient

Convective thermal transport from a surface to a fluid in motion can be related to the heat transfer coefficient, h, the surface-to-fluid temperature difference and the “wetted” surface area, A, in the form

$$q = hA(T_{\text{s}} - T_{\text{fl}}) \tag{19.16}$$

The differences between convection to a rapidly moving fluid, a slowly flowing or stagnant fluid, as well as variations in the convective heat transfer rate among various fluids, are reflected in the values of h. For a particular geometry and flow regime, h may be found from available empirical correlations and/or theoretical relations. Use of Equation 19.6 makes it possible to define the convective thermal resistance, as

$$R_{\text{cv}} \equiv \frac{1}{hA} \tag{19.17}$$

19.2.2.2 Dimensionless Parameters

Common dimensionless quantities that are used in the correlation of heat transfer data are the *Nusselt number*, Nu, which relates the convective heat transfer coefficient to the conduction in the fluid where the subscript, fl, pertains to a fluid property,

$$\mathrm{Nu}_L \equiv \frac{h}{k_{\mathrm{fl}}/L} = \frac{hL}{k_{\mathrm{fl}}} \tag{19.18}$$

the *Prandtl number*, Pr, which is a fluid property parameter relating the diffusion of momentum to the conduction of heat,

$$\mathrm{Pr} \equiv \frac{c_{\mathrm{p}}\mu}{k_{\mathrm{fl}}} \tag{19.19}$$

the *Grashof number*, Gr, which accounts for the bouyancy effect produced by the volumetric expansion of the fluid,

$$\mathrm{Gr}_L \equiv \frac{\rho^2 \beta g L^3 \Delta T}{\mu^2} \tag{19.20}$$

and the *Reynolds number*, Re, which relates the momentum in the flow to the viscous dissipation,

$$\mathrm{Re}_L \equiv \frac{\rho V L}{\mu} \tag{19.21}$$

19.2.2.3 Natural Convection

Despite increasing performance demands and advances in thermal management technology, direct air-cooling of electronic equipment continues to command substantial attention. Natural convection is the quietest, least expensive, and most reliable implementation of direct fluid cooling. In more demanding systems, natural convection cooling with air is often investigated as a baseline design to justify the application of more sophisticated techniques.

In natural convection, fluid motion is induced by density differences resulting from temperature gradients in the fluid. The heat transfer coefficient for this regime can be related to the buoyancy and the thermal properties of the fluid through the *Rayleigh number*, Ra, which is the product of the Grashof and Prandtl numbers,

$$\mathrm{Ra}_L = \frac{\rho^2 \beta g c_{\mathrm{p}}}{\mu k_{\mathrm{fl}}} L^3 \Delta T \tag{19.22}$$

where the fluid properties, ρ, β, c_{p}, μ, and k, are evaluated at the fluid bulk temperature, and ΔT is the temperature difference between the surface and the fluid.

Empirical correlations for the natural convection heat transfer coefficient generally take the form

$$h = C\left(\frac{k_{\mathrm{fl}}}{L}\right)(\mathrm{Ra})^n \tag{19.23}$$

where n is found to be approximately 0.25 for $10^3 < \mathrm{Ra} < 10^9$, representing laminar flow, 0.33 for $10^9 < \mathrm{Ra} < 10^{12}$, the region associated with the transition to turbulent flow, and 0.4 for $\mathrm{Ra} > 10^{12}$, when strong turbulent flow prevails. The precise value of the correlating coefficient, C, depends on the fluid, the geometry of the surface, and the Rayleigh number range. Nevertheless, for natural convection in air from common plate, cylinder, and sphere configurations, it has been found to vary in the relatively narrow range of 0.45–0.65 for laminar flow and 0.11–0.15 for turbulent flow past the heated surface [7].

Vertical channels—Vertical channels formed by parallel PCBs or longitudinal fins are a frequently encountered configuration in natural convection cooling of electronic equipment. The historical work of Elenbaas [8], a milestone of experimental results and empirical correlations, was the first to document a detailed study of natural convection in smooth, isothermal parallel plate channels. In subsequent years, this work was confirmed and expanded both experimentally and numerically by a number of researchers, including Bodoia [9], Sobel et al. [10], Aung [11], Aung et al. [12], Miyatake and Fujii [13], and Miyatake et al. [14].

These studies revealed that the value of the Nusselt number lies between two extremes associated with the separation between the plates or the channel width, b. For wide spacing, the plates appear to have little influence upon one another and the Nusselt number in this case achieves its *isolated plate limit*. On the other hand, for closely spaced plates or for relatively long channels, the fluid attains its *fully developed* velocity profile and the Nusselt number reaches its *fully developed limit*. Intermediate values of the Nusselt number can be obtained from a family of composite expressions developed by Bar-Cohen and Rohsenow [15] and verified by comparison to numerous experimental and numerical studies.

For an isothermal channel, at the fully developed limit, the Nusselt number takes the form

$$\mathrm{Nu}_b = \frac{\mathrm{El}}{C_1} \tag{19.24}$$

where El is the Elenbaas number, defined as

$$\mathrm{El} \equiv \frac{c_\mathrm{p}\rho^2 g\beta(T_\mathrm{w} - T_\mathrm{amb})b^4}{\mu k L} \tag{19.25}$$

where b is the channel spacing, L is the channel length, and $(T_\mathrm{w} - T_\mathrm{amb})$ is the temperature difference between the channel wall and the ambient, or channel inlet. For an asymmetric channel, or one in which one wall is heated and the other is insulated, the appropriate value of C_1 is 12, while for symmetrically heated channels, $C_1 = 24$.

For an isoflux channel, at the fully developed limit, the Nusselt number has been shown to take the form

$$\mathrm{Nu}_b = \sqrt{\frac{\mathrm{El}'}{C_1}} \tag{19.26}$$

where the modified Elenbaas number, El$'$, is defined as

$$\mathrm{El}' \equiv \frac{c_p\rho^2 g\beta q'' b^5}{\mu k^2 x} \tag{19.27}$$

where q'' is the heat flux leaving the channel wall(s). To calculate the maximum wall temperature, the modified Elenbaas number should be based on the channel exit ($x = L$). In this case, the appropriate values of C_1 are 24 and 48 for asymmetric and symmetric heating, respectively. When the channel mid-height ($x = L/2$) wall temperature is of interest, the asymmetric and symmetric C_1 values are 6 and 12, respectively.

In the limit where the channel spacing is very large, the opposing channel walls do not influence each other either hydrodynamically or thermally. This situation may be accurately modeled as heat transfer from an isolated vertical surface in an infinte medium. Natural convection from an isothermal plate can be expressed as

$$\mathrm{Nu}_b = C_2\mathrm{El}^{1/4} \tag{19.28}$$

where McAdams [16] suggests a C_2 value of 0.59. However, more recent research suggests that the available data could be better correlated with $C_2 = 0.515$ [6]. Natural convection from an isoflux plate is typically expressed as

$$\mathrm{Nu}_b = C_2 \mathrm{El}^{1/5} \tag{19.29}$$

with a leading coefficient of 0.631 when based on the maximum ($x = L$) wall temperature and 0.73 when based on the mid-height ($x = L/2$) wall temperature.

Composite Equations—When a function is expected to vary smoothly between two limiting expressions, which are themselves well defined, and when intermediate values are difficult to obtain, an approximate composite relation can be obtained by appropriately summing the two limiting expressions. Using the Churchill and Usagi [17] method, Bar-Cohen and Rohsenow [18] developed composite Nusselt number relations for natural convection in parallel plate channels of the form

$$\mathrm{Nu} = [(\mathrm{Nu}_{\mathrm{fd}})^{-n} + (\mathrm{Nu}_{\mathrm{ip}})^{-n}]^{-1/n} \tag{19.30}$$

where $\mathrm{Nu}_{\mathrm{fd}}$ and $\mathrm{Nu}_{\mathrm{ip}}$ are Nusselt numbers for the fully developed and isolated plate limits, respectively. The correlating exponent n was given a value of 2 to offer good agreement with Elenbaas' [8] experimental results.

For an isothermal channel, combining Equations 19.24 and 19.28 yields a composite relation of the form

$$\mathrm{Nu}_b = \left[\frac{C_3}{\mathrm{El}^2} + \frac{C_4}{\sqrt{\mathrm{El}}}\right]^{-1/2} \tag{19.31}$$

while for an isoflux channel, Equations 19.26 and 19.29 yield a result of the form

$$\mathrm{Nu}_b = \left[\frac{C_3}{\mathrm{El}'} + \frac{C_4}{\mathrm{El}'^{2/5}}\right]^{-1/2} \tag{19.32}$$

Values of the coefficients C_3 and C_4 appropriate to various cases of interest appear in Table 19.1. It is to be noted that the tabulated values reflect a change in the value of the coefficient from 0.59 originally used by by Bar-Cohen and Rohsenow [18] in the isolated, isothermal plate limit to the more appropriate 0.515 value.

In electronic cooling applications where volumetric concerns are not an issue, it is desirable to space PCBs far enough apart that the isolated plate Nusselt number prevails along the surface. In lieu of choosing an infinite plate spacing, the composite Nusselt number may be set equal to 99%, or some other high

TABLE 19.1 Appropriate Values for the C_i Coefficients Appearing in Equations 19.24–19.39

Case	C_1	C_2	C_3	C_4	C_5	C_6	C_7
Isothermal							
Symmetric heating	24	0.515	576	3.77	4.43	0.00655	2.60
Asymmetric heating	12	0.515	144	3.77	3.51	0.0262	2.06
Isoflux							
Symmetric heating							
Maximum temp.	48	0.63	48	2.52	9.79	0.105	2.12
Midheight temp.	12	0.73	12	1.88	6.80	0.313	1.47
Asymmetric heating							
Maximum temp.	24	0.63	24	2.52	7.77	0.210	1.68
Midheight temp.	6	0.73	6	1.88	5.40	0.626	1.17

fraction of its associated isolated plate value. The composite Nusselt number relation may then be solved for the appropriate channel spacing.

For an isothermal channel, the channel spacing that maximizes the rate of heat transfer from individual PCBs takes the form

$$b_{\max} = \frac{C_5}{P^{1/4}} \tag{19.33}$$

where

$$P = \frac{c_p \rho^2 g\beta(T_{\mathrm{w}} - T_{\mathrm{amb}})}{\mu k L} = \frac{\mathrm{El}}{b^4}$$

while for an isoflux channel, the channel spacing that minimizes the PCB temperature for a given heat flux takes the form

$$b_{\max} = \frac{C_5}{R^{1/5}} \tag{19.34}$$

where

$$R = \frac{c_p \rho^2 g\beta q''}{\mu k^2 L} = \frac{\mathrm{El}'}{b^5}$$

Values of the coefficient C_5 appropriate to various cases of interest appear in Table 19.1.

Optimum Spacing—In addition to being used to predict heat transfer coefficients, the composite relations presented may be used to optimize the spacing between plates. For isothermal arrays, the optimum spacing maximizes the total heat transfer from a given base area or the volume assigned to an array of plates or PCBs. In the case of isoflux parallel plate arrays, the total array heat transfer for a given base area may be maximized by increasing the number of plates indefinitely, though the plate will experience a commensurate increase in temperature. Thus, it is more appropriate to define the optimum channel spacing for an array of isoflux plates as the spacing that will yield the maximum volumetric heat dissipation rate per unit temperature difference. Despite this distinction, the optimum spacing is found in the same manner.

The total heat transfer rate from an array of vertical, single-sided plates can be written as

$$\frac{Q_{\mathrm{T}}}{LsWk\Delta T} = \left(\frac{\mathrm{Nu}}{b(b+d)}\right) \tag{19.35}$$

where the number of plates, $m = W/(b+d)$, d is the plate thickness, W is the width of the entire array, and s is the depth of the channel. The optimum spacing may be found by substituting the appropriate composite Nusselt number equation into Equation 19.35, taking the derivative of the resulting expression with respect to b, and setting the result equal to zero. Use of an isothermal composite Nusselt number in Equation 19.35 yields a relation of the form

$$(2b + 3d - C_6 P^{3/2} b^7)_{\mathrm{opt}} = 0 \tag{19.36}$$

or

$$b_{\mathrm{opt}} = \frac{C_7}{P^{1/4}} \tag{19.37}$$

when d, the plate thickness, is negligible. Use of an isoflux composite Nusselt number yields

$$(b + 3d - C_6 R^{3/5} b^4)_{\text{opt}} = 0 \tag{19.38}$$

or

$$b_{\text{opt}} = \frac{C_7}{R^{1/5}} \quad (d = 0) \tag{19.39}$$

Values of the coefficients C_6 and C_7 appropriate to various cases of interest appear in Table 19.1.

Limitations—These smooth-plate relations have proven useful in a wide variety of applications and have been shown to yield very good agreement with measured empirical results for heat transfer from arrays of PCBs. However, when applied to closely spaced PCBs, where the spacing is of the order of the component height, these equations tend to underpredict heat transfer in the channel due to the presence of between-package "wall flow" and the nonsmooth nature of the channel surfaces [19].

19.2.2.4 Forced Convection

For forced flow in long or very narrow parallel-plate channels, the heat transfer coefficient attains an asymptotic value (a fully developed limit), which for symmetrically heated channel surfaces is equal approximately to

$$h = \frac{4k_{\text{fl}}}{d_{\text{e}}} \tag{19.40}$$

where d_{e} is the *hydraulic diameter* defined in terms of the flow area, A, and the wetted perimeter of the channel, P_{w}

$$d_{\text{e}} \equiv \frac{4A}{P_{\text{w}}}$$

In the inlet zones of such parallel-plate channels and along isolated plates, the heat transfer coefficient varies with the distance from the leading edge. The low-velocity, or laminar flow, average convective heat transfer coefficient for $\text{Re} \leq 2 \times 10^5$ is given by [7]

$$h = 0.664 \left(\frac{k}{L}\right) \text{Re}^{1/2}\text{Pr}^{1/3} \tag{19.41}$$

where k is the fluid thermal conductivity and L is the characteristic dimension of the surface. This heat transfer coefficient decreases asymptotically toward the fully developed value given by Equation 19.40.

A similar relation applies to flow in tubes, pipes, ducts, channels, and annuli with the equivalent diameter, d_{e} serving as the characteristic dimension in both the Nusselt and Reynolds numbers. For laminar flow, $\text{Re} \leq 2100$

$$\frac{\bar{h} d_{\text{e}}}{k} = 1.86 \left[\text{RePr}\left(\frac{d_{\text{e}}}{L}\right)\right]^{1/3} \left(\frac{\mu}{\mu_{\text{w}}}\right)^{0.14} \tag{19.42}$$

which is attributed to Sieder and Tate [20] and where μ_{w} is the viscosity of the convective medium at the wall temperature. Observe that Equations 19.41 and 19.42 show that the heat transfer coefficient from the surface to the fluid is highest for short channels and decreases as L increases.

In higher velocity turbulent flow, the dependence of the convective heat transfer coefficient on the Reynolds number increases and, in the range $\mathrm{Re} > 3 \times 10^5$, is typically given by [7]

$$h = 0.036\left(\frac{k}{L}\right)(\mathrm{Re})^{0.80}(\mathrm{Pr})^{1/3} \tag{19.43}$$

In pipes, tubes, channels, ducts, and annuli, transition to turbulent flow occurs at an equivalent diameter-based Reynolds number of approximately 10,000. Thus, the flow regime bracketed by $2100 \leq \mathrm{Re} \leq 10,000$ is usually referred to as the transition region. Hausen [21] has provided the correlation

$$\frac{hd_\mathrm{e}}{k} = 0.116[\mathrm{Re} - 125](Pr)^{1/3}\left(1 + \frac{d_\mathrm{e}}{L}\right)^{2/3}\left(\frac{\mu}{\mu_\mathrm{w}}\right) \tag{19.44}$$

and Sieder and Tate [20] give for turbulent flow

$$\frac{hd_\mathrm{e}}{k} = 0.23(\mathrm{Re})^{0.80}(\mathrm{Pr})^{1/3}\left(\frac{\mu}{\mu_\mathrm{w}}\right) \tag{19.45}$$

Additional correlations for the coefficient of heat transfer in forced convection for various configurations may be found in the heat transfer textbooks [6,22–24].

19.2.3 Phase Change Heat Transfer

When heat exchange is accompanied by evaporation of a liquid or condensation of a vapor, the resulting flow of vapor toward or away from the heat transfer surface and the high rates of thermal transport associated with the latent heat of the fluid can provide significantly higher heat transfer rates than single-phase heat transfer alone.

19.2.3.1 Boiling

Boiling heat transfer displays a complex dependence on the temperature difference between the heated surface and the saturation temperature (boiling point) of the liquid. In nucleate boiling, the primary region of interest, the ebullient heat transfer rate is typically expressed in the form of the Rohsenow [25] equation

$$q = \mu_\mathrm{f} h_\mathrm{fg}\sqrt{\frac{g(\rho_\mathrm{f} - \rho_\mathrm{g})}{\sigma}}\left[\frac{c_\mathrm{pf}}{C_\mathrm{sf} Pr_\mathrm{f}^{1.7} h_\mathrm{fg}}\right]^{1/r}(T_\mathrm{s} - T_\mathrm{sat})^{1/r} \tag{19.46}$$

where $1/r$ is typically correlated with a value of 3, and C_sf is a function of characteristics of the surface/fluid combination. Rohsenow recommended that the fluid properties in Equation 19.46 be evaluated at the liquid saturation temperature.

For pool boiling of the dielectric liquid FC-72 ($T_\mathrm{sat} = 57°\mathrm{C}$ at 101.3 kPa) on a plastic-pin-grid-array (PPGA) chip package, Watwe et al. [26] obtained values of 7.47 for $1/r$ and 0.0075 for C_sf. At a surface heat flux of 10 W/cm^2, the wall superheat at 101.3 kPa is nearly 30°C, corresponding to an average surface temperature of approximately 86°C.

The departure from nucleate boiling, or "Critical Heat Flux" (CHF), places an upper limit on the use of the highly efficient boilng heat transfer mechanism. CHF can be significantly influenced by system parameters such as pressure, subcooling, heater thickness and properties, and dissolved gas content. Watwe et al. [26] presented the following equation to predict the pool boiling critical heat flux of dielectric coolants from microelectronic components and under a variety of parametric conditions.

$$\mathrm{CHF} = \left\{\frac{\pi}{24} h_{\mathrm{fg}} \sqrt{\rho_{\mathrm{g}}} [\sigma_{\mathrm{f}} g(\rho_{\mathrm{f}} - \rho_{\mathrm{g}})]^{1/4}\right\} \left(\frac{\delta\sqrt{\rho_{\mathrm{h}} c_{\mathrm{ph}} k_{\mathrm{h}}}}{\delta\sqrt{\rho_{\mathrm{h}} c_{\mathrm{ph}} k_{\mathrm{h}}} + 0.1}\right)$$

$$\times \{1 + [0.3014 - 0.01507L'(P)]\} \left\{1 + 0.03 \left[\left(\frac{\rho_{\mathrm{f}}}{\rho_{\mathrm{g}}}\right)^{0.75} \frac{c_{\mathrm{pf}}}{h_{\mathrm{fg}}}\right] \Delta T_{\mathrm{sub}}\right\} \tag{19.47}$$

The first term on the right-hand side of Equation 19.47 is the classical Kutateladze–Zuber prediction, which is in the upper limit on the saturation value of CHF on very large horizontal heaters. The second term represents the effects of heater thickness and thermal properties on the critical heat flux. The third term in Equation 19.47 accounts for the influence of the heater size, where

$$L' = L\sqrt{\frac{g(\rho_{\mathrm{f}} - \rho_{\mathrm{g}})}{\sigma_{\mathrm{f}}}} \tag{19.48}$$

This third term is only to be included when its value is larger than unity (i.e., $0.3014 - 0.01507L' > 0$) as small heaters show an increase in CHF over large heaters. The last term is an equation representing the best-fit line through the experimental data of Watwe et al. [26] and represents the influence of subcooling on CHF. The pressure effect on CHF is embodied in the Kutateladze–Zuber and subcooling model predictions, which make up Equation 19.47, via the thermophysical properties. Thus, Equation 19.47 can be used to estimate the combined influences of various system and heater parameters on CHF. The critical heat flux, under saturation conditions at atmospheric pressure, for a typical dielectric coolant like FC-72 and for a 1 cm component is approximately 15 W/cm^2. Alternatively, at 2 atm and 30°C of subcooling CHF for FC-72 could be expected to reach 22 W/cm^2.

19.2.3.2 Condensation

Closed systems involving an evaporative process must also include some capability for vapor condensation. Gerstmann and Griffth [27] correlated film condensation on a downward-facing flat plate as

$$\mathrm{Nu} = 0.81\,\mathrm{Ra}^{0.193} \quad 10^{10} > \mathrm{Ra} > 10^{8} \tag{19.49}$$

$$\mathrm{Nu} = 0.69\,\mathrm{Ra}^{0.20} \quad 10^{8} > \mathrm{Ra} > 10^{6} \tag{19.50}$$

where

$$\mathrm{Nu} \equiv \frac{h}{k}\left(\frac{\sigma}{g(\rho_{\mathrm{f}} - \rho_{\mathrm{g}})}\right)^{1/2} \tag{19.51}$$

$$\mathrm{Ra} \equiv \frac{g\rho_{\mathrm{f}}(\rho_{\mathrm{f}} - \rho_{\mathrm{g}})h_{\mathrm{fg}}}{k\mu\Delta T}\left(\frac{\sigma}{g(\rho_{\mathrm{f}} - \rho_{\mathrm{g}})}\right)^{3/2} \tag{19.52}$$

The Nusselt number for laminar film condensation on vertical surfaces was correlated by Nusselt [28] and later modified by Sadasivan and Lienhard [29] as

$$\mathrm{Nu} = \frac{hL}{k_{\mathrm{f}}} = 0.943\left[\frac{g\Delta\rho_{\mathrm{fg}} L^3 h'_{\mathrm{fg}}}{k_{\mathrm{f}} \nu_{\mathrm{f}} (T_{\mathrm{sat}} - T_{\mathrm{c}})}\right]^{1/4} \tag{19.53}$$

where

$$h'_{fg} = h_{fg}(1 + C_c Ja)$$

$$C_c = 0.683 - \frac{0.228}{Pr_l}$$

$$Ja = \frac{c_{pf}(T_{sat} - T_c)}{h_{fg}}$$

19.2.3.3 Phase Change Materials

In recent years there has been growing use of solid–liquid phase change materials (PCM) to help mitigate the deleterious effects of transient "spikes" in the power dissipation and/or environmental load imposed on RF modules. This is of particular importance for outdoor modules, where PCMs can serve to smooth diurnal variations in the air temperature and solar radiations. To determine the mass of PCM needed to absorb a specified thermal load at a constant (melting) temperature, it is necessary to obtain the latent heat of fusion of that material and insert it in the following relation:

$$m = \frac{Q}{h_{fs}} \tag{19.54}$$

19.2.4 Flow Resistance

The transfer of heat to a flowing gas or liquid that is not undergoing a phase change results in an increase in the coolant temperature from an inlet temperature of T_{in} to an outlet temperature of T_{out}, according to

$$T_{out} - T_{in} = \frac{q}{\dot{m}c_p} \tag{19.55}$$

On the basis of this relation, it is possible to define an effective flow resistance, R_{fl}, as

$$R_{fl} \equiv \frac{1}{\dot{m}c_p} \tag{19.56}$$

where $\dot{m}$, the mass flow rate, is given in kg/s.

In multicomponent systems, determination of individual component temperatures requires knowledge of the fluid temperature adjacent to the component. The rise in fluid temperature relative to the inlet value can be expressed in a flow thermal resistance, as done in Equation 19.56. When the coolant flow path traverses many individual components, care must be taken to use R_{fl} with the total heat absorbed by the coolant along its path, rather than the heat dissipated by an individual component. For system-level calculations aimed at determining the average component temperature, it is common to base the flow resistance on the average rise in fluid temperature, that is, one-half the value indicated by Equation 19.55.

19.2.5 Radiative Heat Transfer

Unlike conduction and convection, radiative heat transfer between two surfaces or between a surface and its surroundings is not linearly dependent on the temperature difference and is expressed instead as

$$q = \sigma A \mathcal{F}(T_1^4 - T_2^4) \tag{19.57}$$

where $\mathcal{F}$ includes the effects of surface properties and geometry and σ is the Stefan–Boltzman constant, $\sigma = 5.67 \times 10^{-8}$ W/m^2 K^4. For modest temperature differences, this equation can be linearized to the form

$$q = h_r A(T_1 - T_2) \tag{19.58}$$

where h_r is the effective "radiation" heat transfer coefficient

$$h_r = \sigma \mathcal{F}(T_1^2 + T_2^2)(T_1 + T_2) \tag{19.59}$$

and, for small $\Delta T = T_1 - T_2$, h_r is approximately equal to

$$h_r = 4\sigma \mathcal{F}(T_1 T_2)^{3/2} \tag{19.60}$$

where T_1 and T_2 must be expressed in absolute degrees Kelvin. It is of interest to note that for temperature differences of the order of 10 K with absolute temperatures around room temperature, the radiative heat transfer coefficient, h_r for an ideal (or "black") surface in an absorbing environment is approximately equal to the heat transfer coefficient in natural convection of air.

Noting the form of Equation 19.58, the radiation thermal resistance, analogous to the convective resistance, is seen to equal

$$R_r \equiv \frac{1}{h_r A} \tag{19.61}$$

19.2.6 Environmental Heat Transfer

In applying the foregoing thermal transport relations to microwave equipment located outdoors, attention must be devoted to properly characterizing the atmospheric conditions and including both incoming solar radiation and outgoing night-sky radiation in the heat balance relations. While best results will be obtained by using precise environmental specifications for the "microclimate" at the relevant location, more general specifications may be of use in early stages of product design. The external environment can vary in temperature from −50°C to +50°C, representing the polar regions at one extreme and the subtropical deserts at the other, and experience a range in air pressure from 76 kPa (11 psi), at high plateaus, to 107 kPa (15.5 psi), in deep rift valleys. Incident solar fluxes at noon can reach 1 kW/m^2 on a horizontal surface, but more typically may average 0.5 kW/m^2, of both direct and diffuse radiation, during the peak solar hours. The outgoing long-wave radiation from an outdoor module exposed to the clear nighttime sky falls in the range of 0.01–0.1 kW/m^2 [30]. It may be anticipated that convective heat transfer coefficients on large exposed surfaces at sea level will attain values of 6 W/m^{2}K for still air and 75 W/m^{2}K at wind velocities approaching 100 km/h. To determine the surface temperature of an outdoor module, use can be made of the heat balance relation equating the incoming heat—from the microwave components and solar load—with the outgoing heat—by radiation and convection, as

$$q_{rf} + q_{solar} = q_{rad} + q_{conv} \tag{19.62}$$

or

$$\begin{aligned} q_{rf} &= q_{rad} + q_{conv} - q_{solar} \\ &= \sigma A_{surf} \mathcal{F}(T_{surf}^4 - T_{sky}^4) + h_{conv} A_{surf}(T_{surf} - T_{amb}) - \alpha A_{surf} S \end{aligned} \tag{19.63}$$

or

$$T_{surf} = \frac{(q_{rf}/A_{surf}) + \alpha S}{\sigma \mathcal{F}(T_{surf}^2 + T_{sky}^2)(T_{surf} + T_{sky}) + h_{conv}} + T_{sky} \tag{19.64}$$

where S is the solar incidence (W/m^2) and T_{sky} is the effective sky temperature (K) (typically equal to ambient temperature during the day and up to 20 K below the air temperature on a dry, clear night).

19.3 Thermal Resistance Networks

The expression of the governing heat transfer relations in the form of thermal resistances greatly simplifies the first-order thermal analysis of electronic systems. Following the established rules for resistance networks, thermal resistances that occur sequentially along a thermal path can be simply summed to establish the overall thermal resistance for that path. In similar fashion, the reciprocal of the effective overall resistance of several parallel heat transfer paths can be found by summing the reciprocals of the individual resistances. In refining the thermal design of an electronic system, prime attention should be devoted to reducing the largest resistances along a specified thermal path and/or providing parallel paths for heat removal from a critical area.

While the thermal resistances associated with various paths and thermal transport mechanisms constitute the "building blocks" in performing a detailed thermal analysis, they have also found widespread application as "figures-of-merit" in evaluating and comparing the thermal efficacy of various packaging techniques and thermal management strategies.

19.3.1 Chip Module Thermal Resistance

19.3.1.1 Definition

The thermal performance of alternative chip packaging techniques is commonly compared on the basis of the overall (junction-to-coolant or junction-to-ambient) thermal resistance, R_{ja} This packaging figure-of-merit is generally defined in a purely empirical fashion,

$$R_{ja} \equiv \frac{T_j - T_{fl}}{q_c} \tag{19.65}$$

where T_j and T_{fl} are the junction and coolant (fluid) temperatures, respectively, and q_c is the chip heat dissipation.

Unfortunately, however, most measurement techniques are incapable of detecting the actual junction temperature, that is, the temperature of the small volume at the interface of p-type and n-type semiconductors. Hence, this term generally refers to the average temperature or a representative temperature on the chip.

Examination of various packaging techniques reveals that the junction-to-coolant thermal resistance is, in fact, composed of an internal, largely conductive resistance and an external, primarily convective resistance. As shown in Figure 19.6, the internal resistance, R_{jc} is encountered in the flow of dissipated heat from the active chip surface through the materials used to support and bond the chip and on to the case of the integrated circuit package. The flow of heat from the case directly to the coolant, or indirectly through a fin structure and then to the coolant, must overcome the external resistance, R_{ex}.

19.3.1.2 Internal Thermal Resistance

As discussed previously, conductive thermal transport is governed by the Fourier equation, which can be used to define a conduction thermal resistance, as in Equation 19.3. In flowing from the chip to the package surface or case, the heat encounters a series of resistances associated with individual layers of materials, starting with the chip (silicon, galium arsenide, indium phosphide, etc.) and continuing thru solder, copper, alumina, and epoxy, as well as the contact resistances that occur at the interfaces between pairs of materials. Although the actual heat flow paths within a chip package are rather complex and may shift to accommodate varying external cooling situations, it is possible to obtain a first-order estimate of the internal resistance by assuming that power is dissipated uniformly across the chip surface

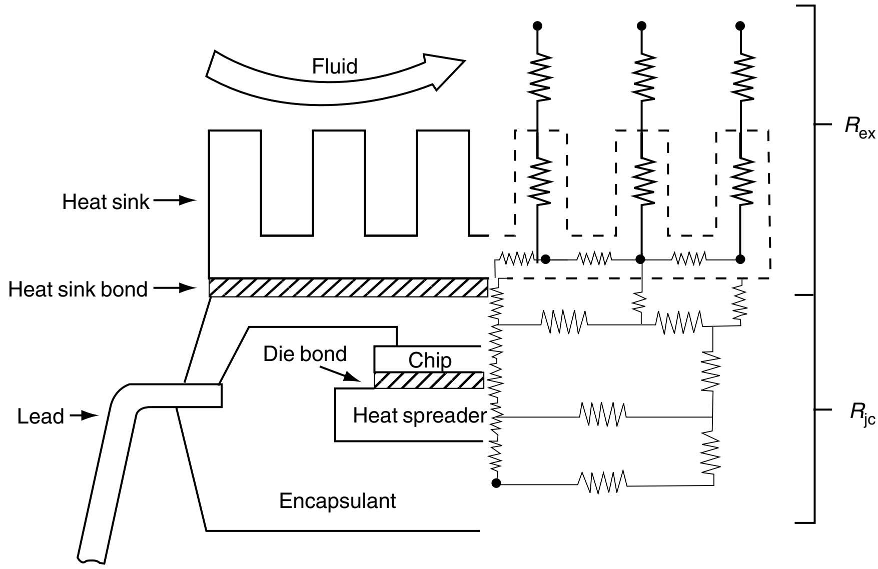


FIGURE 19.6 Primary thermal resistances in a single chip package. (From A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995. With permission.)

and that heat flow is largely one-dimensional. To the accuracy of these assumptions, the following equation:

$$R_{jc} = \frac{T_j - T_c}{q_c} = \sum \frac{\Delta x}{kA} \tag{19.66}$$

can be used to determine the internal chip module resistance, where the summed terms represent the conduction thermal resistances posed by the individual layers, each with thickness Δx. As the thickness of each layer decreases and/or the thermal conductivity and cross-sectional area increase, the resistance of the individual layers decreases. Values of R_{cd} for packaging materials with typical dimensions can be found via Equation 19.66 or Figure 19.7, to range from 2 K/W for a 1000 mm^2 by 1 mm-thick layer of epoxy encapsulant to 0.0006 K/W for a 100 mm^2 by 25 μm (1 mil) thick layer of copper. Similarly, the values of conduction resistance for typical "soft" bonding materials are found to lie in the range of approximately 0.1 K/W for solders and 1–3 K/W for epoxies and thermal pastes for typical $\Delta x/A$ ratios of 0.25–1.0.

Comparison of theoretical and experimental values of R_{jc} reveals that the resistances associated with compliant, low-thermal conductivity bonding materials and the spreading resistances, as well as the contact resistances at the lightly loaded interfaces within the package, often dominate the internal thermal resistance of the chip package. It is, thus, not only necessary to correctly determine the bond resistance but also to add the values of R_{sp}, obtained from Equation 19.7 and/or Figure 19.4, and R_{co} from Equations 19.9 or 19.12 to the junction-to-case resistance calculated from Equation 19.66. Unfortunately, the absence of detailed information on the voidage in the die-bonding and heat-sink attach layers and the present inability to determine, with precision, the contact pressure at the relevant interfaces, conspire to limit the accuracy of this calculation.

19.3.1.3 External Resistance

An application of Equations 19.41 or 19.43 to the transfer of heat from the case of a chip module to the coolant shows that the external resistance, $R_{ex} = 1/hA$, is inversely proportional to the wetted

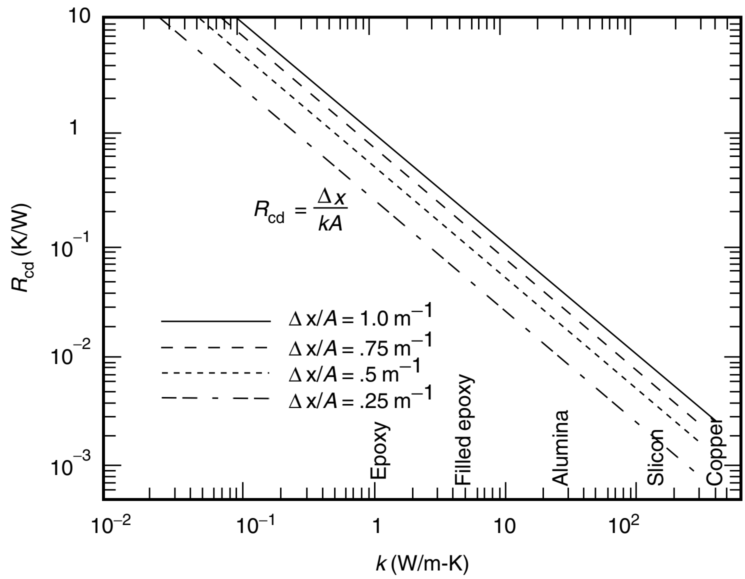


FIGURE 19.7 Conductive thermal resistances for packaging materials. (From A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995. With permission.)

surface area and to the coolant velocity to the 0.5 to 0.8 power and directly proportional to the length scale in the flow direction to the 0.5 to 0.2 power. It may, thus, be observed that the external resistance can be strongly influenced by the fluid velocity and package dimensions and that these factors must be addressed in any meaningful evaluation of the external thermal resistances offered by various packaging technologies.

Values of the external resistance for a variety of coolants and heat transfer mechanisms are shown in Figure 19.8 for a typical component wetted area of 10 cm^2 and a velocity range of 2–8 m/s. They are seen to vary from a nominal 100 K/W for natural convection in air, to 33 K/W for forced convection in air, to 1 K/W in fluorocarbon liquid forced convection, and to less than 0.5 K/W for boiling in fluorocarbon liquids. Clearly, larger chip packages will experience proportionately lower external resistances than the displayed values. Moreover, conduction of heat through the leads and package base into the PCB or substrate will serve to further reduce the effective thermal resistance.

In the event that the direct cooling of the package surface is inadequate to maintain the desired chip temperature, it is common to attach finned heat sinks, or compact heat exchangers, to the chip package. These heat sinks can considerably increase the wetted surface area, but may act to reduce the convective heat transfer coefficient by obstructing the flow channel. Similarly, the attachment of a heat sink to the package can be expected to introduce additional conductive resistances, in the adhesive used to bond the heat sink and in the body of the heat sink. Typical air-cooled heat sinks can reduce the external resistance to approximately 10–15 K/W in natural convection and to as low as 3–5 K/W for moderate forced convection velocities.

When a heat sink or compact heat exchanger is attached to the package, the external resistance accounting for the bond-layer conduction and the total resistance of the heat sink, R_{hs}, can be expressed as

$$R_{ex} = \frac{T_c - T_{fl}}{q_c} = \sum \left(\frac{\Delta x}{kA}\right)_b + R_{hs} \tag{19.67}$$

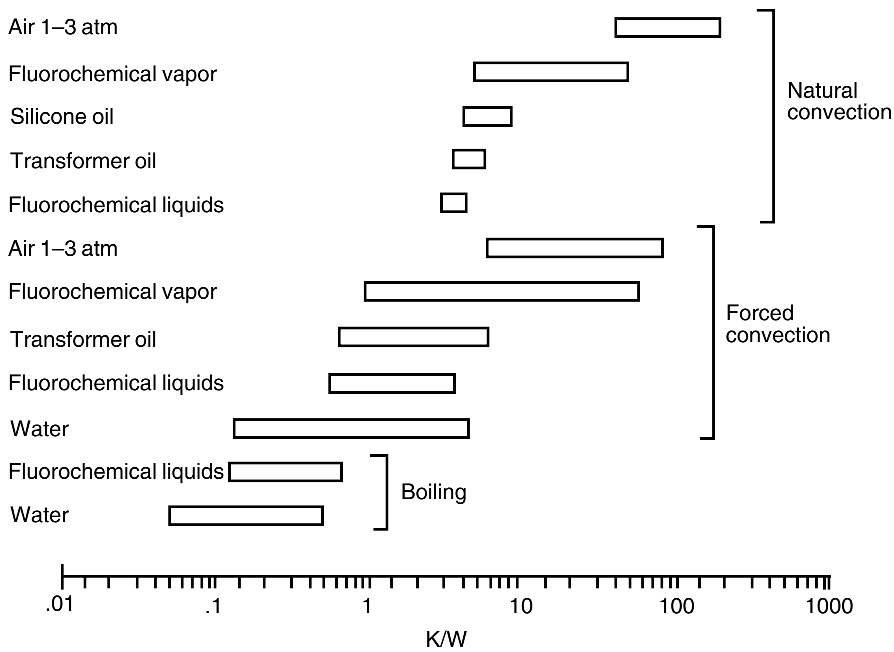


FIGURE 19.8 Typical external (convective) thermal resistances for various coolants and cooling modes. (From A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995. With permission.)

where R_{hs}

$$R_{\text{hs}} = \left[\frac{1}{nhA_{\text{f}}\eta} + \frac{1}{h_{\text{b}}A_{\text{b}}}\right]^{-1} \tag{19.68}$$

is the the parallel combination of the resistance of the n fins

$$R_{\text{f}} = \frac{1}{nhA_{\text{f}}\eta} \tag{19.69}$$

and the *bare* or base surface not occupied by the fins

$$R_{\text{b}} = \frac{1}{h_{\text{b}}A_{\text{b}}} \tag{19.70}$$

Here, the base surface area is $A_{\text{b}} = A - A_{\text{f}}$, and use of the heat transfer coefficient, h_{b}, is meant to recognize that the heat transfer coefficient that is applied to the base surfaces is not necessarily equal to that applied to the fins.

An alternative expression for R_{hs} involves an *overall surface efficiency*, η_{o}, defined by

$$\eta_{\text{o}} = 1 - \frac{nA_{\text{f}}}{A}(1 - \eta) \tag{19.71}$$

where A is the total surface composed of the base surface and the finned surfaces of n fins

$$A = A_{\text{b}} + nA_{\text{f}} \tag{19.72}$$

In this case, it is presumed that $h_b = h$ so that

$$R_{hs} = \frac{1}{h\eta_o A} \tag{19.73}$$

In an optimally designed fin structure, η can be expected to fall in the range of 0.50–0.70 [31]. Relatively thick fins in a low velocity flow of gas are likely to yield fin efficiencies approaching unity. This same unity value would be appropriate, as well, for an unfinned surface and, thus, serves to generalize the use of Equation 19.67 to all package configurations.

19.3.1.4 Total Resistance—Single Chip Packages

To the accuracy of the assumptions employed in the preceding development, the overall single chip package resistance, relating the chip temperature to the inlet temperature of the coolant, can be found by summing the internal, external, and flow resistances to yield

$$\begin{aligned} R_{ja} &= R_{jc} + R_{ex} + R_{fl} \\ &= \sum \frac{\Delta x}{kA} + R_{int} + R_{sp} \\ &= \frac{1}{\eta hA} + \left(\frac{Q}{q}\right)\left(\frac{1}{2\rho Q c_p}\right) \end{aligned} \tag{19.74}$$

In evaluating the thermal resistance by this relationship, care must be taken to determine the effective cross-sectional area for heat flow at each layer in the module and to consider possible voidage in any solder and adhesive layers.

As previously noted in the development of the relationships for the external and internal resistances, Equation 19.74 shows R_{ja} to be a strong function of the convective heat transfer coefficient, the flowing heat capacity of the coolant, and geometric parameters (thickness and cross-sectional area of each layer). Thus, the introduction of a superior coolant, use of thermal enhancement techniques that increase the local heat transfer coefficient, or selection of a heat transfer mode with inherently high heat transfer coefficients (e.g., boiling) will all be reflected in appropriately lower external and total thermal resistances. Similarly, improvements in the thermal conductivity and reduction in the thickness of the relatively low conductivity bonding materials (such as soft solder, epoxy, or silicone) would act to reduce the internal and total thermal resistances.

19.3.1.5 Weighted-Average Modification of R_{jc}

The commonly used junction-to-case thermal resistance, relying on just a single case temperature, can be used with confidence only when the package case is nearly isothermal. In a more typical packaging configuration, when substantial temperature variations are encountered among and along the external surfaces of the package [32–34] the use of the reported R_{jc} can lead to erroneous chip temperature predictions. This is especially of concern in the analysis and design of plastic chip packages, owing to the inherently high thermal resistance of the plastic encapsulant and the package anisotropies introduced by the large differences in the conductivity between the lead frame and/or heat spreader and the plastic encapsulant. Since R_{jc} is strictly valid only for an isothermal package surface, a method must be found to address the individual contributions of the various surface segments according to their influence on the junction temperature.

Following Krueger and Bar-Cohen [35], it is convenient to introduce the expanded R_{jc} methodology with a thermal model of a chip package that can be approximated by a network of three thermal resistances connected in parallel, from the chip to the top, sides, and bottom of the package, respectively. This type of compact model is commonly referred to as a "star network" and, in this model, the heat flow from

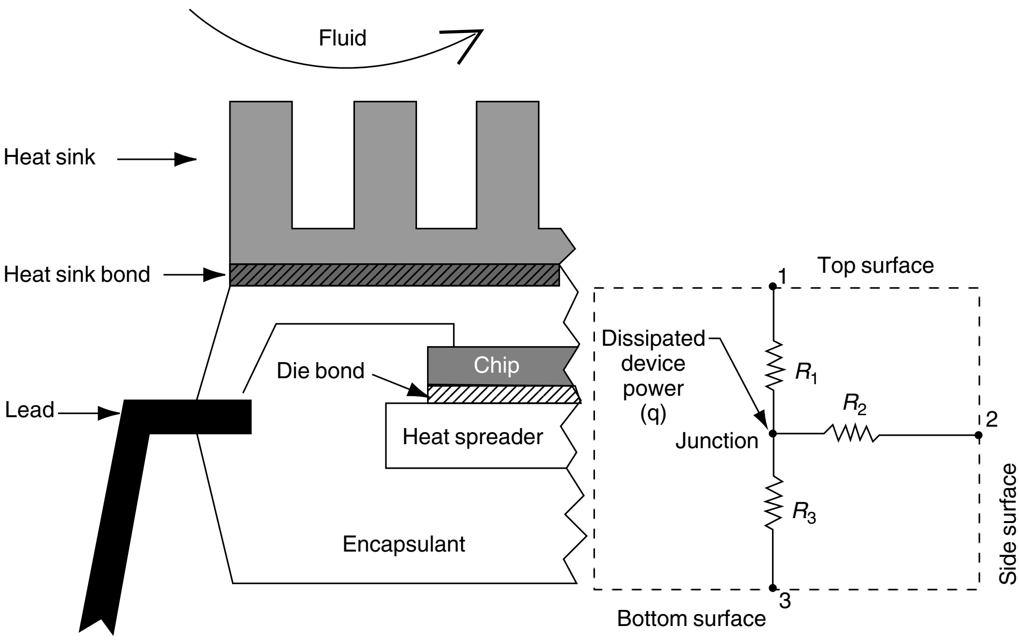


FIGURE 19.9 Geometry of a 28-lead PLCC device. (a) The compact model schematic and (b) the actual device cross-section. (From W.B. Krueger and A. Bar-Cohen, *IEEE CHMT Trans.*, 15(5), 691–698, 1992. With permission.)

the chip is

$$q = q_1 + q_2 + q_3 \tag{19.75}$$

or

$$q = \frac{T_j - T_1}{R_1} + \frac{T_j - T_2}{R_2} + \frac{T_j - T_3}{R_3} \tag{19.76}$$

This compact model of an electronic device is shown schematically in Figure 19.9.

Equation 19.76 can be rearranged to yield the dependence of the chip (or junction) temperature on the temperature of the three surface segments as

$$T_j = \left(\frac{R_2 R_3}{R_s}\right) T_1 + \left(\frac{R_3 R_1}{R_s}\right) T_2 + \left(\frac{R_1 R_2}{R_s}\right) T_3 + \left(\frac{R_1 R_2 R_3}{R_s}\right) q \tag{19.77}$$

where $R_s = R_1 R_2 + R_1 R_3 + R_2 R_3$.

Equation 19.77 may be generalized to admit n-distinct elements along the package surface, or

$$T_j = \sum_{k=1}^{n} I_k T_k + I_{n+1} q \tag{19.78}$$

A comparison of Equations 19.77 and 19.78 shows that the coefficients of the specified surface temperatures, the I_ks, are totally determined by the internal resistances of the chip package

$$I_1 = \frac{R_2 R_3}{R_s} \quad I_2 = \frac{R_3 R_1}{R_s}$$
$$I_3 = \frac{R_1 R_2}{R_s} \quad I_4 = \frac{R_1 R_2 R_3}{R_s} \tag{19.79}$$

The temperature coefficients needed to generate a junction temperature relation of the form shown in Equation 19.78 can thus be determined from previously calculated internal resistances or, in the absence of such values, by extraction from empirical data or numerical results for the junction temperature. Furthermore, it is to be noted that the sum of the coefficients of the various surface temperatures is identically equal to unity and that the power dissipation coefficient, $I_{n+1}q$, is, in fact, the familiar R_{jc}, the isothermal, junction-to-case thermal resistance. Consequently, Equation 19.78 may be rewritten as

$$T_{\text{j}} = \sum_{k=1}^{n} I_k T_k + R_{\text{jc}} q \tag{19.80}$$

or, returning to R_{jc}

$$R_{\text{jc}} = T_{\text{j}} - \frac{\sum_{k=1}^{n} I_k T_k}{q} = \frac{T_{\text{j}} - \bar{T}_{\text{c}}}{q} \tag{19.81}$$

where $\bar{T}_{\text{c}}$ is the average case temperature

$$\bar{T}_{\text{c}} = \frac{\sum_{k=1}^{n} A_k}{A_{\text{T}}} T_k \tag{19.82}$$

where A_k is the surface area of the kth surface and A_{T} is the surface area of the entire package.

In many applications, chip packages are cooled selectively along particular exposed surfaces. One such example is a package cooled from the top and side surfaces while the bottom surface is insulated. Thermally active surfaces may vary from application to application, and the thermal analyst needs to quantify the effect of thermally insulating one or more areas on a package of known thermal resistance. For the assumptions used in the development of the expanded R_{jc} model, insulation of surface-m results in zero heat flow through resistance, R_m. This causes the temperature of surface-m to equal the chip temperature. With this in mind, the junction temperature for a package with a single insulated surface given by Equation 19.80 is found to equal

$$T_{\text{j}} = \sum_{k \neq m} \left(\frac{I_k}{1 - I_m} \right) T_k + \left(\frac{I_{n+1}}{1 - I_m} \right) q \tag{19.83}$$

The weighted average case temperature for this thermal configuration is found to equal

$$\bar{T}_{\text{c}} = \sum_{k \neq m} \left(\frac{I_{\text{jc}}}{1 - I_m} \right) T_k \tag{19.84}$$

and the modified junction to case resistance, R^*_{jc}, is

$$R^*_{\text{jc}} = \frac{R_{\text{jc}}}{1 - I_m} \tag{19.85}$$

19.3.2 Multichip Modules

The thermostructural complexity of multichip modules in current use hampers effective thermal characterization and introduces significant uncertainty in any attempt to compare the thermal performance of these packaging configurations. Variations in heat generation patterns across the active chips (reflecting differences in functional requirements and the duty cycle among the macrocells constituting a particular chip), as well as nonuniformities in heat dissipation among the chips assembled in a single module, further complicate this task. While nonthermal issues (e.g., electromagnetic crosstalk) may often be the dominant limiting factors in RF multichip modules, the high power associated with microwave devices make it essential that the thermal performance of this packaging configuration be analyzed and reported in a consistent manner.

To establish a common basis for comparison of multichip modules, it is possible to neglect the on-chip and chip-to-chip variations and consider that the heat generated by each chip flows through a unit cell of the module structure to the external coolant [36,37]. For a given structure, increasing the area of the unit cell allows heat to spread from the chip to a larger cross section, reducing the heat flux at some of the thermally critical interfaces and at the convectively cooled surfaces. Consequently, the thermal performance of a multichip module can be best represented by the area-specific thermal resistance, that is, the temperature difference between the chip and the coolant divided by the substrate heat flux, expressed in units of $K/(W/cm^2)$. This figure of merit is equivalent to the inverse of the overall heat transfer coefficient, U, commonly used in the compact heat exchanger literature.

Despite significant variation in design and fabrication, the leading edge water-cooled and air-cooled modules of the late 1980s provided a specific thermal resistance of approximately 20°C for every watt per square centimeter at the substrate. A decade later, the thermally best multichip modules of the 1990s offered specific thermal resistance values between 5 and 10 $K/(W/cm^2)$. Increasingly high chip-level heat fluxes coupled with increasing silicon efficiency (ratio of chip to substrate surface area) are expected to drive target-specific thermal resistances toward 1 $K/(W/cm^2)$ by the end of this decade. In addition, owing to the surging demand for portable computing/communication devices of the 2000s, traditional MCM configurations are giving way to integrated RFICs, homogeneous and heterogeneous chip stacks, and other system-in-package (SiP) technologies. These significantly more complex packages (many 3-D) will require highly sophisticated compact models and figures of merit for appropriate thermal characterization.

19.3.3 Radar System Applications

The frequent demand for high radiated power in civilian and military radar systems, for ranging and detection of remote objects, has led to the development of many high-energy microwave systems. Owing to the inefficiencies inherent in electrical-to-microwave energy conversion and the management of RF energy, the operation of such radar equipment often results in significant heat dissipation. To avoid catastrophic failures, to achieve the reliability targets of the system, and to satisfy the frequency stability requirements of the RF tubes, system requirements commonly specify temperature control to within several degrees Celsius around 150°C. These thermal requirements necessitate the use of aggressive thermal management techniques, including the use of high flow rate liquid forced convection, pool and flow boiling, and high pressure drop, air-cooled compact heat exchangers [7].

In traditional, mechanically steered radar systems, much of the total power dissipation occurs in the power tubes (e.g., klystron, gyrotrons, amplitrons), with secondary dissipation in the rotary joints and wave guides. Heat release along the collector surfaces of the tubes is proportional to the RF power of the system and can range from 1 kW to as much as 20 kW, with a peak local flux of several W/cm^2 to, perhaps, several thousand W/cm^2, at operating temperatures of 150°C [7]. Similar though less severe thermal problems are encountered in RF rotary joints and waveguides, where heat fluxes of 5–10 W/cm^2, with allowable temperatures of 150°C, may be encountered.

Growing applications of active antenna array elements, utilizing amplifier diodes and solid-state phase shifters to provide electronic steering of the radiated RF beam, have introduced new dimensions into the thermal control of microwave equipment. In such "phased array" radar systems, high power tubes and waveguides can be eliminated with low power RF delivered to each antenna element. The low conversion efficiency of the amplifier diodes results in diode heat dissipation between 1 and 10 W and local heat fluxes comparable to power transistors. In "phased array" antennae, precise shaping of the radiated wave often requires relatively cool ($<90°C$) ferrite phase shifters and no significant temperature variations ($<10°C$) across the elements used in the array.

References

1. M. Pecht, P. Lall, and E. Hakim, The influence of temperature on integrated circuit failure mechanisms, In *Advances in Thermal Modeling of Electronic Components and Systems*, A. Bar-Cohen and A.D. Kraus (Eds.), New York: ASME Press, vol. 3, 1992.
2. R.A. Morrison, Improved avionics reliability through phase change conductive cooling, *Proceedings of the IEEE National Telesystems Conference*, Galveston, TX, pp. B5.6.1–B5.6.5, 1982.
3. K.J. Negus, M.M. Yovanovich, and J. V. Beck, On the non-dimensionalization of constriction resistance for semi-infinite heat flux tubes, *J. Heat Trans.*, 111(3), 804–807, 1989.
4. M.M. Yovanovich and V.W. Antonetti, Applications of thermal contact resistance theory to electronic packages, In *Advances in Thermal Modeling of Electronic Components and Systems*, A. Bar-Cohen and A.D. Kraus (eds.), New York: Hemisphere Publishing Co., 79–128, 1988.
5. M.M. Yovanovich, personal communication, 1990.
6. F.P. Incropera and D.P. Dewitt, *Introduction to Heat Transfer*, New York: John Wiley & Sons, 1996.
7. A.D. Kraus and A. Bar-Cohen, *Thermal Analysis and Control of Electronic Equipment*, New York: McGraw-Hill, 1983.
8. W. Elenbaas, Heat dissipation of parallel plates by free convection, *Physica*, 9(1), 665–671, 1942.
9. J.R. Bodia and J.F. Osterle, The development of free convection between heated vertical plates, *J. Heat Transfer*, 84, 40–44, 1964.
10. N. Sobel, F. Landis, and W.K. Mueller, Natural convection heat transfer in short vertical channels including the effect of stagger, *Proceedings of the Third International Heat Transfer Conference, Chicago*, 2, pp. 121–125, 1966.
11. W. Aung, Fully developed laminar free convection between vertical plates heated asymmetrically, *Int. J. Heat Mass Transfer*, 15, 40–44, 1972.
12. W. Aung, L.S. Fletcher, and V. Sernas, Developing laminar free convection between vertical flat plates with asymmetric heating, *Int. J. Heat Mass Transfer*, 15, 2293–2308, 1972.
13. O. Miyatake and T. Fujii, Free convection heat transfer between vertical parallel plates—one plate isothermally heated and the other thermally insulated, *Heat Transfer Jpn. Res.*, 3, 30–38, 1972.
14. O. Miyatake, T. Fujii, M. Fujii, and H. Tanaka, Natural convection heat transfer between vertical parallel plates—one plate with a uniform heat flux and the other thermally insulated, *Heat Transfer Jpn. Res.*, 4, 25–33, 1973.
15. A. Bar-Cohen and W.M. Rohsenow, Thermal optimum spacing of vertical, natural convection cooled, vertical plates, *J. Heat Trans.*, 106, 116–122, 1984.
16. W.H. McAdams, *Heat Transmission*, New York: McGraw-Hill, 1954.
17. S.W. Churchill and R. Usagi, A general expression for the correlation of rates of transfer and other phenomena, *AIChE J.*, 18(6), 1121–1138, 1972.
18. A. Bar-Cohen and W.M. Rohsenow, Thermally optimum spacing of vertical, natural convection cooled, parallel plates, In *Heat Transfer in Electronic Equipment, ASME HTD 20, ASME WAM*, Washington, D.C., 11–18, 1981.
19. A. Bar-Cohen, Bounding relations for natural convection heat transfer from vertical printed circuit boards, *Proceedings of the IEEE*, September 1985.
20. E.N. Sieder and G.E. Tate, Heat transfer and pressure drops of liquids in tubes, *Ind. Eng. Chem.*, 28, 1429–1435, 1936.
21. H. Hausen, Darstelling des wärmeuberganges in rohren durch verallgemeinerte potenzbeziehungen, *VDI Z*, 4, 91–98, 1943.
22. A. Bejan, *Heat Transfer*, New York: John Wiley & Sons, 1993.
23. J.P. Holman, *Heat Transfer*, New York: McGraw-Hill, 1990.
24. J.H. Lienhard, *A Heat Transfer Textbook*, Englewood Cliffs, NJ: Prentice-Hall, 1987.

25. W.M. Rohsenow, A method of correlating heat transfer data for surface boiling of liquids, *Transactions of ASME (reprinted in 3rd ASME/JSME Thermal Engineering Joint Conference, 1, 503-512)*, 74, 969–976, 1951.
26. A.A. Watwe, A. Bar-Cohen, and A. McNeil, Combined pressure and subcooling effects on pool boiling from a PPGA chip package, *ASME J. Electronics Packaging*, March 1997.
27. F. Gertsmann and P. Griffth, Laminar film condensation on the underside of horizontal and inclined surfaces, *Int. J. Heat Mass Transfer*, 10, 567–580, 1966.
28. W.Z. Nusselt, Die oberflächencondensation der wasserdamfes, *Z. Ver. Deut. Ing.*, 60, 541–569, 1916.
29. P. Sadasivan and J.H. Lienhard, Sensible heat correction in laminar film boiling and condensation, *J. Heat Transfer*, 109, 545–546, 1987.
30. C. Rambach and A. Bar-Cohen, Nocturnal heat rejection by skyward radiation, In *Future Energy Production Systems*, J.C. Denton and N. Afgan (eds.), NY: Hemisphere Publishing, pp. 713–726, 1976.
31. A.D. Kraus and A. Bar-Cohen, *Design and Analysis of Heat Sinks*, New York: John Wiley & Sons, 1995.
32. J.A. Andrews, Package thermal resistance model dependency on equipment design, *IEEE CHMT Trans.*, 11(4), 528–537, 1988.
33. S.S. Furkay, Thermal characterization of plastic and surface mount components, *IEEE CHMT Trans.*, 11(4), 521-527, 1988.
34. E.A. Wilson, Factors influencing the interdependence of R_{jc} and R_{ca}, *Proceedings of the Technical Program, 1st Annual Conference—International Electronic Packaging Society*, Cleveland, OH, pp. 247–255, 1981.
35. W.B. Krueger and A. Bar-Cohen, Thermal characterization of a PLCC–expanded R_jc methodology, *IEEE CHMT Trans.*, 15(5), 691–698, 1992.
36. A. Bar-Cohen, Thermal management of air- and liquid-cooled multichip modules, *IEEE CHMT Trans.*, CHMT-10(2), 159–175, 1987.
37. A. Bar-Cohen, Addendum and correction to thermal management of air- and liquid-cooled multichip modules, *IEEE CHMT Trans.*, CHMT-11(3), 333–334, 1988.

20

Electronic Hardware Reliability

Diganta Das
Michael Pecht
CALCE-University of Maryland

20.1 Introduction

Reliability is the ability of a product to properly function, within specified performance limits, for a specified period of time, under the life cycle application conditions. Achieving product reliability over the complete life cycle demands an approach that consists of a set of tasks, each requiring total engineering and management commitment and enforcement. These tasks impact electronic hardware reliability through the selection of materials, structural geometries and design tolerances, manufacturing processes and tolerances, assembly techniques, shipping and handling methods, operational conditions, and maintenance and maintainability guidelines. The tasks are as follows:

- Define realistic product requirements and constraints determined by factors such as required operating and storage life, the life cycle application profile, performance requirements, environmental regulations, disposal requirements, size, weight, and cost. The product requirements should be defined based on both the customer's needs and the manufacturer's capability to meet those needs.
- Define the product life cycle environment by specifying all relevant assembly, storage, handling, shipping, operating, maintenance, and disposal conditions for the fielded product.
- Select the parts and materials required for the product, using a well-defined assessment procedure that ensures that the parts selected meet quality* requirements, are capable of delivering the expected performance and reliability in the application, and will be available to sustain the product throughout its life cycle. The materials and parts must be characterized for variabilities in properties (e.g., mechanical, thermo-physical, and electrical). Knowledge of these variabilities is required to assess design margins.
- Identify the potential failure sites and failure mechanisms† by which the product can be expected to fail. Potential architectural and stress interactions must also be defined and assessed in this identification process. Appropriate measures must be implemented to assure control of those failure mechanisms through design updates and usage management.
- Design to the usage profile of the product and process capability of the manufacturing process (i.e., the quality level that can be controlled in manufacturing and assembly), considering the potential failure sites and failure mechanisms. The design stress spectra, the part test spectra, and the qualification and acceptance test spectra must be based on the anticipated life cycle usage conditions. The proposed product must survive the life cycle environment, while considering manufacturability, quality, reliability, and cost-effectiveness, and be available to the market in a timely manner.
- Qualify the product manufacturing and assembly processes. Key process characteristics in all the manufacturing and assembly processes required to make the part must be identified, measured, and optimized. Tests should be conducted to verify the results for complex products. The goal of this step is to provide a physics-of-failure (PoF) basis for design decisions, with an assessment of all possible failure mechanisms for the anticipated product. If all the processes are in control and the design is valid, then product testing is not warranted and is therefore not cost-effective. This represents a transition from product test, analysis, and screening to process test, analysis, and screening.
- Monitor and control the manufacturing and assembly processes addressed in the design, so that process shifts do not arise. Each process may involve screens and tests to assess statistical process control.
- Monitor the life cycle usage of the product using closed-loop management procedures. This includes realistic inspection and maintenance procedures.

The approach of design for reliability is based on the PoF methodology [1–5]. This chapter describes the methodologies to undertake each task. PoF reliability prediction methods are based on root cause analyses of underlying electronic component failure processes and are used to proactively predict distributions of time to failure (TTF). This is achieved by developing failure models through experimental studies of the critical materials and by developing computer-assisted stress models of the intended product. The merits

*Quality is a measure of a part's ability to meet the workmanship criteria of the manufacturer.

†A failure mechanism is a process (such as creep, fatigue, or wear) through which a defect nucleates and grows as a function of stresses (such as thermal, mechanical, electromagnetic, or chemical loadings) ultimately resulting in the degradation or failure of a product.

of PoF approaches over traditional statistical empirical reliability assurance approaches are discussed by Dasgupta [5].

20.2 Product Requirements and Constraints

Product development is a process in which the perceived need for a product leads to the definition of requirements, which are translated into a design. A product is realized when a process is developed to manufacture a product that meets the established requirements. The product definition is directly derived from the needs and constraints of the market. Product realization requires technical expertise and engineering ingenuity. The final success in marketing any product is based on the ability to deliver products that meet current market requirements in terms of size, weight, shape, portability, functionality and performance, ergonomics, aesthetics, safety, and cost [6,7].

One of the first steps of product development is the process of transforming broad goals and vague concepts into realizable, concrete requirements. The company's core competencies, culture, goals, and customers define the requirements and constraints for product designs. Figure 20.1 shows that the product definition results from a combination of marketing and business-driven product requirements, design and manufacturing constraints, and other external influences. Marketing often takes the lead in determining the product's requirements and constraints. For products (parts or assemblies) that must be fabricated as part of a specific larger product (e.g., an ignition module that is attached to an automobile engine), the requirements and constraints are usually defined by the customer (the manufacturer of the product or system that the subproduct fits into), and the marketing function is less involved.

The development of product specifications begins with the presentation of an initial set of requirements and constraints. The initial requirements are formulated into a preliminary requirements document, which should be approved by people ranging from engineers to corporate management to customers (the actual people involved in the approval depends on the company and the product). Once the requirements are approved, engineering prepares a specification indicating the exact set of requirements that are practical to implement.

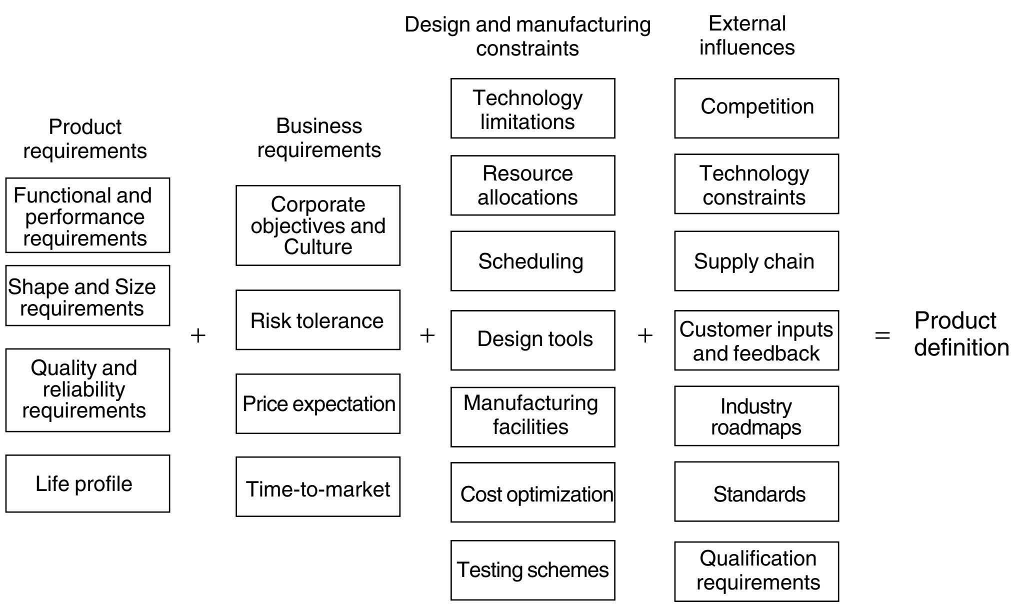


FIGURE 20.1 Constraints in the product definition process.

Design decisions are a balance of all the requirements, as per the final specification of the product. The design may be adjusted to reduce cost or to improve attributes such as ergonomics, safety, performance, quality, and reliability.

From the perspective of requirements and constraints, products can be classified into three types: multicustomer products, single-customer products, and custom products—with variability of the targeted customer population. To make reliable products, there should be a joint effort between the supplier and the customer. The IEEE Reliability Program Standard—1332 [8] presents the relationship between the supplier and customer, in terms of reliability objectives. The standard identifies three reliability objectives:

- The supplier, working with the customer, shall determine and understand the customer's requirements and product needs so that a comprehensive design specification can be generated.
- The supplier shall structure and follow a series of engineering activities such that the resulting product satisfies the customer's requirements and product needs with regard to product reliability.
- The supplier shall include activities that assure the customer that the reliability requirements and product needs have been satisfied.

If the product is for direct sale to end users, marketing usually takes the lead in defining the product's requirements and constraints through interaction with the customer's marketplace, examination of product sales figures, and analysis of the competition. Alternatively, if the product is a subsystem that fits within a larger product, the requirements and constraints are determined by the customer's product into which the subsystem fits. The product definition process involves multiple influences and considerations. Figure 20.1 shows a diagram of constraints in the product definition process.

Two prevalent risks in requirements and constraints definition are inclusion of irrelevant requirements and omission of relevant requirements. The inclusion of irrelevant requirements can lead to unnecessary expenditures and time for design and testing. Irrelevant or erroneous requirements result from two sources: requirements created by persons who do not understand the constraints and opportunities implicit in the product definition and inclusion of requirements for historical reasons. The latter are requirements that get "cut and pasted" into requirements documents from a previous product. No one knows exactly why the requirement is included, but no one is brave enough to remove it, simply because no one takes the time to see the obvious (mistake) or to question the norm. The omission of critical requirements may cause the product to be dysfunctional or may significantly reduce the effectiveness and appeal of the product by not meeting the necessary attributes. This may reduce the market size of the product and delay the product launch, thus shrinking the time window for getting return on investment.

The actual content of the requirements document is application-specific. The requirements document may also assign a priority to each requirement. When engineers read the requirements document and decide what subset of the requirements can be accommodated in the product, they need to know which requirements are the most important. Schedule and cost may not be included in the requirements document for some products if the requirements document is released to other (internal or external) organizations for bids. A single person cannot realistically define all product requirements for a product. To make the requirements realistic and useful, people from different disciplines within a company should contribute to, review, and approve the product requirements, depending on the specific product, the customers involved, and the corporate culture. For example, for human-safety-critical products, legal representatives should attend the approval to identify legal considerations with respect to the implementation of the parts selection and management team's directives. The results of capturing product requirements and constraints allow the design team to choose parts and develop products that conform to company objectives.

Once a set of requirements has been completed, the product engineering function creates a response to the requirements in the form of a specification. The specification states the requirements that must be met; the schedule for meeting the requirements; the identification of those persons who will perform the work; and the identification of potential risks. Differences in the requirements document and the preliminary specification become the topic of trade-off analyses.

Once product requirements are defined and the design process begins, the process of continuously comparing the product's requirements to the actual product design begins. As the product's design becomes increasingly detailed, through selection of parts and choice of implementation strategies, it becomes increasingly important to track the product's characteristics (e.g., size, weight, performance, functionality, reliability, and cost) in relation to the original product requirements. The rationale for making changes should be documented and approved. The completeness with which requirements' tracking is performed can significantly reduce future product redesign costs. Planned redesigns or design refreshes through technology monitoring and use of roadmaps ensure that a company is able to market new products or redesigned versions of old products in a timely, effective manner to retain its customer base and ensure continued profits.

20.3 The Product Life Cycle Environment

The life cycle environment of the product plays a significant role in determining the product requirements and part selection. It influences the product design and development decisions, qualification and specification processes, reliability assessments, quality assurance, product safety, warranty and support commitments, and regulatory conformance.

The life cycle of a product consists of manufacturing and assembly, testing, rework, storage, transportation and handling, operation, repair and maintenance, and disposal. The life cycle loads include thermal (steady-state temperature, temperature ranges, temperature cycles, temperature gradients), mechanical (pressure levels, pressure gradients, vibrations, shock loads, acoustic levels), chemical (aggressive or inert environments, ozone, pollution humidity levels, contamination, fuel spills), physical (radiation, electromagnetic interference, altitude), and/or operational loading conditions (power, power surge, heat dissipation, current, voltage spikes). These loads, either individually or in various combinations, may influence the reliability of the product. The extent and rate of product degradation depend on the nature, magnitude, and duration of exposure to such loads.

The life cycle environmental loads need to be quantified in terms of range of possible values and expected variability of these values. Ideally, the designer may want to know the distribution of the loads experienced by the product. Several methods for quantifying the life cycle loads are discussed in the following sections. Although methods such as conducting *in situ* monitoring provide more accurate information, they are not always employed due to the involving nature of the monitoring process. Designers may usually resort to easier (and often less accurate) estimates developed from market studies and field trials. Effective quantification of environmental loads may require the use of all or a combination of the methods discussed in the later sections.

20.3.1 Market Studies and Standards-Based Profiles as Sources of Data

Market surveys and other reports generated independently by some agencies* or developed by industries as a part of their design process are used as the basis for environmental load characterization. These kinds of data are derived most often from a remotely similar kind of environment and give a very coarse estimate of the actual environmental loads, which the targeted product will experience. Standards-based profiles should only be applied after considerable similarity analysis.

20.3.2 *In Situ* Monitoring of Environmental Loads

Environmental and usage loads experienced by the product in its life cycle can be monitored *in situ*. These data are often collected using sensors, either mounted externally or integrated with the product and supported by telemetry systems. Devices such as health and usage monitoring systems are popular in aircraft and helicopters for *in situ* monitoring of usage and environmental loads. Load distributions should be developed from data obtained by monitoring products used by different customers, ideally from

*These agencies include focus groups in organizations and standards committees like those who develop military standards. For example, IPC SM-785 specifies the use and extreme temperature conditions for electronic products categorized under different industry sectors such as telecommunication, commercial, military, and so forth.

various geographical locations where the product is used. The data should be collected over a sufficient period to provide an accurate estimate of the loads and their variation over time. *In situ* monitoring has the potential to provide the most accurate account of load history for use in health (condition) assessment and design of future products [9].

20.3.3 Field Trial Records, Service Records, and Failure Records

Field trial records are also sometimes used to get estimates on the environmental profiles. They provide estimates on the environmental profiles experienced by the product. The data depend on the duration and conditions of the trials, and can be extrapolated to get an estimate of actual environmental conditions. Service records and failure records usually document the causes for scheduled or unscheduled maintenance and the nature of failure in the product, which might have been due to certain environmental or usage conditions. These conditions are sometimes used to estimate the kinds of environments the product might be subjected to.

20.3.4 Data on Load Histories of Similar Parts, Assemblies, or Products

Similarity analysis is a technique for estimating environmental loads when sufficient field histories for similar products are available. Before using data on existing products for proposed designs, the characteristic differences in design and application for the two products need to be reviewed. Changes and discrepancies in the conditions of the two products should be critically analyzed to ensure the applicability of available loading data for the new product. For example, electronics inside a washing machine in a commercial laundry is expected to experience a wider distribution of loads and use conditions (due to several users) and higher usage rates compared with a home washing machine. These differences should be considered during similarity analysis.

20.4 Parts Selection and Management

Almost all products include subsystems that are built by other organizations. It is incumbent upon the system integrators to select the parts (materials) that have sufficient quality and are capable of delivering the expected performance and reliability in the application. Owing to changes in technology trends, the evolution of complex supply–chain interactions and new market challenges, shifts in consumer demand, and continuing standards reorganization, a cost-effective and efficient parts selection, and management process is needed to perform this assessment. The goal is to provide an “eyes-on, hands-off” approach to parts selection and management that enables companies to organize and conduct fact-finding processes to select parts with improved quality, integrity, application-specific reliability, and cost-effectiveness, and make an informed companywide decision about parts selection and management based on company resources, policies, culture, and goals and customer demands.

The parts selection and management process is usually not carried out by a single individual, but rather by a multidisciplinary team. The parts management team, as a whole, is responsible for all the issues associated with the effect of a part on a system and the effect of the system on a part. This group performs communication and organizations functions such as assigning parts selection and management responsibilities to groups within the company, establishing communication channels within and outside the company, and managing information flow within the team and to departments outside the team. It also develops and implements the technical tasks like identifying process and assessment criteria, and acceptability levels, and applying the parts selection and management methodology to the candidate part. When necessary, it helps in identifying potential supplier intervention procedures, authorizing such action when required, and ensuring the associated effectiveness. All through this implementation and communication, the group is also required to periodically monitor the process and results and make improvements to the methodology continuously.

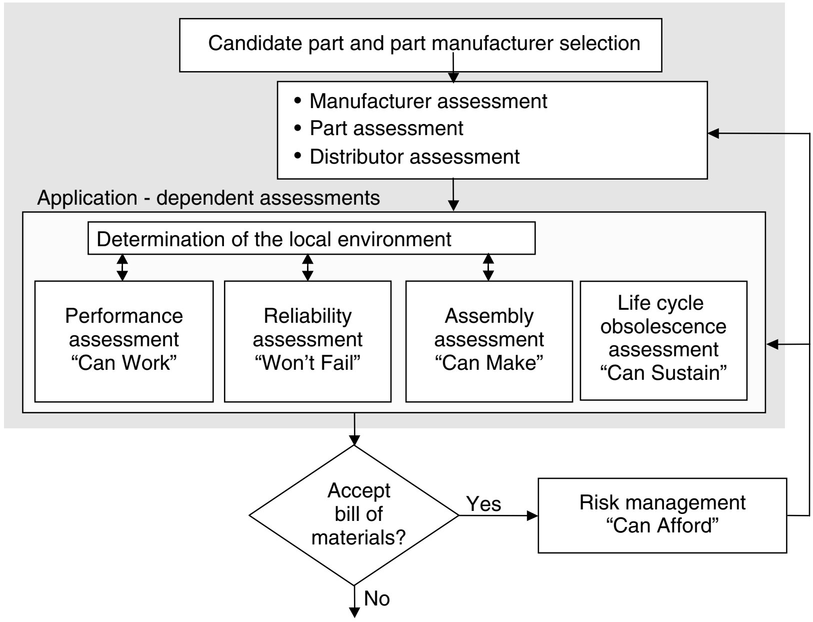


FIGURE 20.2 Parts selection and management methodology.

The methodology shown in Figure 20.2 describes this approach to parts selection and management. The overall purpose is to help organizations to maximize profits, to provide product differentiation, to effectively utilize the global supply chain, and to assess, mitigate, and manage the life cycle risks in selecting and using parts. Several of these assessment methods directly impact product reliability and those steps are described in the following sections. Several other steps such as assembly assessment and life cycle obsolescence assessment are also performed in this process. While those steps impact overall product performance and profitability, their impact on reliability are not direct.

20.4.1 Manufacturer, Part, and Distributor Assessment

In the manufacturer assessment, the part manufacturer's ability to produce parts with consistent quality is evaluated. In the part assessment, the candidate part's quality and reliability is gauged. The distributor assessment evaluates the distributor's ability to provide parts without affecting the initial quality and reliability, and to provide certain specific services such as part problem and change notifications. The product manufacturer's parts selection and management team defines the minimum acceptability criteria for this assessment based on its requirements. If the part satisfies the minimum acceptability criteria, the candidate part then moves to "application-dependent assessments."

If the part is found unacceptable due to nonconformance with the minimum acceptability criteria, some form of equipment supplier intervention is considered [6,7]. If equipment supplier intervention is not feasible due to economic or schedule considerations, the candidate part may be rejected. If, however, equipment supplier intervention is considered necessary, then the intervention action items are identified and their cost and schedule implications are analyzed through the "risk management" process step.

20.4.2 Performance Assessment

The goal of performance assessment is to evaluate the part's ability to meet the performance requirements (e.g., functional, mechanical, and electrical) of the product. In order to increase performance, products may incorporate features that could make them less reliable than proven, lower-performance products. Increasing the number of parts, although improving performance also increases product complexity, can lead to lower reliability unless compensating measures are taken [10]. In such situations, product reliability can be maintained only if part reliability is increased or part redundancy is built into the product. Each of these alternatives, in turn, must be assessed against the incurred cost.

In general, there are no distinct boundaries for parameters such as mechanical load, voltage, current, temperature, and power dissipation, above which immediate failure will occur and below which a part will operate indefinitely [9]. However, there is often a minimum and a maximum limit beyond which the part will not function properly, or at which the increased complexity required will not offer an advantage in cost-effectiveness. Part manufacturers' ratings or users' procurement ratings are generally used to determine these limiting values. Equipment manufacturers who integrate such parts into their products need to adapt their design so that the parts do not experience conditions beyond their absolute maximum ratings, even under the worst possible operating conditions (e.g., supply voltage variations, load variations, and signal variations) [11]. It is the responsibility of the parts selection and management team to establish that the electrical, mechanical, or functional performance of the part is suitable for the life cycle conditions of the particular product. If a product must be operated outside the manufacturer-specified operating conditions, then uprating* may have to be considered [12].

20.4.3 Reliability Assessment

Reliability assessment results provide information about the ability of a part to meet the required performance specifications in its life cycle application environment for a specified period. If the parametric and functional requirements of the system cannot be met within the required local environment, then the local environment may have to be modified, or a different part may have to be used.

Reliability assessment is conducted through the use of integrity test data, virtual qualification results, or accelerated test results. Integrity is a measure of the appropriateness of the tests conducted by the manufacturer and of the part's ability to survive those tests. Integrity monitoring tests are conducted by the part manufacturer to monitor part and process changes, and the ongoing material or process changes specific to the part. Integrity test data (often available from the part manufacturer) are examined in light of the life cycle conditions and applicable failure mechanisms and models. If the magnitude and duration of the life cycle conditions are less severe than those of the integrity tests, and if the test sample size and results are acceptable, the part reliability is acceptable. If the integrity test data are insufficient to validate part reliability in the application, then virtual qualification should be considered.

Virtual qualification is a simulation-based methodology used to identify the dominant failure mechanisms associated with the part under the life cycle conditions, to determine the acceleration factor for a given set of accelerated test parameters, and to determine the TTFs corresponding to the identified failure mechanisms. Virtual qualification allows the operator to optimize the part parameters (e.g., dimensions, materials) so that the minimum TTF of any part is greater than the expected product life. If virtual qualification proves insufficient to validate part reliability, accelerated testing should be performed. Once the appropriate test procedures, conditions, and sample sizes are determined, accelerated testing can be conducted by the part manufacturer, the equipment supplier, or third-party test facilities. Accelerated testing results are used to predict the life of a product in its field application by computing an acceleration factor that correlates the accelerated test conditions and the actual field conditions. Whether integrity test data, virtual qualification results, accelerated test results, or a combination thereof are used, each applicable failure mechanism to which the part is susceptible must be addressed. If part reliability is not ensured through the reliability assessment process, the equipment supplier must consider an alternate part or product redesign. If redesign is not considered a viable option, the part should be rejected and an alternate part must be selected. If the part must be used in the application, redesign options may include load (e.g., thermal, electrical, and mechanical) management techniques, vibration and shock,

*The term uprating was coined by Michael Pecht to distinguish it from *upscreening*, which is a term used to describe the practice of attempting to create a part equivalent to a higher quality by additional screening of a part (e.g., screening a JANTXV part to JANS requirements).

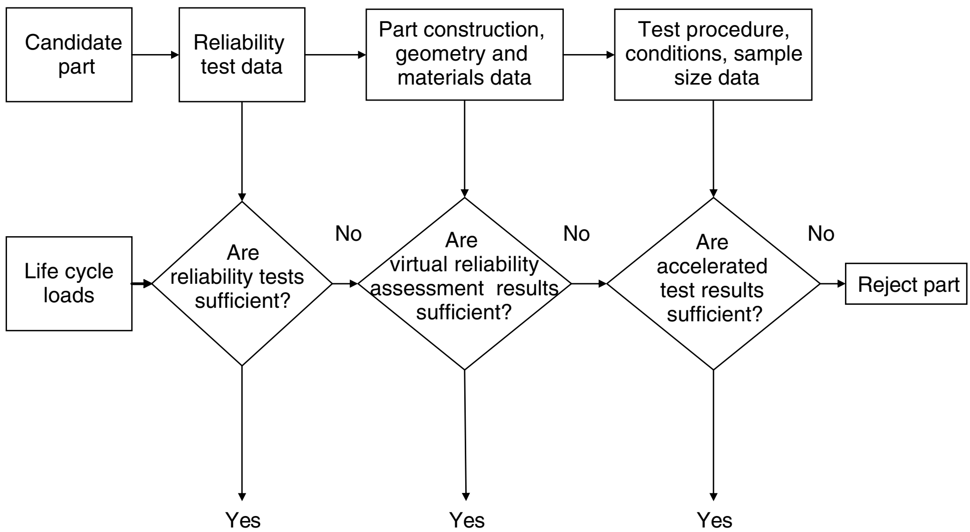


FIGURE 20.3 Decision process for part reliability assessment.

damping, and modifying assembly parameters. If product design changes are made, part reliability must be reassessed. This decision process is illustrated in the diagram in Figure 20.3.

20.4.4 Risk Management

After a part is accepted, resources must be applied to managing the life cycle of the part, including supply chain management, obsolescence assessment, manufacturing and assembly feedback, manufacturer warranties management, and field failure and root-cause analysis. Including time, data, opportunity, and money, resources determine whether risks should be managed or not. The risks associated with including a part in the product fall into two categories:

- *Managed risks*: risks that the product development team chooses to proactively manage by creating a management plan and performing a prescribed regimen of monitoring the part's field performance, manufacturer, and manufacturability
- *Unmanaged risks*: risks that the product development team chooses not to proactively manage

If risk management is considered necessary, a plan should be prepared. The plan should contain details about how the part is monitored (data collection), and how the results of the monitoring feedback into various parts selection and management processes. The feasibility, effort, and cost involved in management processes prior to the final decision to select the part must be considered. Feedback regarding the part's assembly performance, field performance, and sales history is essential to ascertain the risks.

20.5 Failure Modes, Mechanisms, and Effects Analysis

Electronic hardware is typically a combination of board, components, and interconnects, all with various failure mechanisms by which they can fail in the life cycle environment. A potential failure mode is the manner in which a failure can occur—that is, the ways in which the item fails to perform its intended design function, or performs the function but fails to meet its objectives [13,14]. A failure mode can also be defined as any physically observable change caused by a failure mechanism. Failure modes are closely related to the functional and performance requirements of the product. Failure mechanisms are the processes by which a specific combination of physical, electrical, chemical, and mechanical stresses induces failures.

The competitive marketplace places demands on manufacturers to identify cost-effective methods of improving the product development process. In particular, the industry has been interested in an approach of understanding potential product failures that might affect product performance over time. The failure mode and effects analysis (FMEA) methodology is a procedure to recognize and evaluate the potential failure of a product and its effects, and to identify actions that could eliminate or reduce the likelihood of the potential failure to occur [14,15]. Many organizations within the electronics industry have employed or required the use of FMEA; but in general, this methodology has not provided satisfaction, except for the purpose of safety analysis [16].

A limitation of the FMEA procedure is that it does not identify the product failure mechanisms and models in the analysis and reporting process. Investigation of the possible failure modes and mechanisms of the product aids in developing failure-free and reliable designs. A design team must be aware of the possible failure mechanisms to design hardware capable of withstanding loads without failing. Failure mechanisms and their related physical models are also important for planning tests and screens to audit nominal design and manufacturing specifications, as well as the level of defects introduced by excessive variability in manufacturing and material parameters.

Failure modes, mechanisms, and effects analysis (FMMEA) is a novel methodology that has been developed to address weaknesses in the traditional FMEA process [16]. The purpose of FMMEA is to identify potential failure mechanisms and models for all potential failures modes and to prioritize failure mechanisms. FMMEA enhances the value of the traditional FMEA and failure mode, effects, and criticality analysis (FMECA) by identifying high-priority failure mechanisms in order to create an action plan to mitigate their effects. High-priority failure mechanisms determine the operational stresses and the environmental and operational parameters that need to be controlled. Models for the failure mechanisms help in the design and development of the product.

FMMEA is based on understanding the relationships between product requirements and the physical characteristics of the product (and their variation in the production process), the interactions of product materials with loads (stresses at application conditions) and their influence on the product susceptibility to failure with respect to the use conditions. This involves finding the failure mechanisms and reliability models to quantitatively evaluate the susceptibility to failure. In addition to the information gathered and used for FMEA, FMMEA uses life cycle environmental and operating conditions and the duration of the intended application with knowledge of the active stresses and potential failure mechanisms. The FMMEA process is summarized in Figure 20.4 and is described in the following sections.

20.5.1 System Definition, Elements, and Functions

As illustrated in Figure 20.4, a FMMEA process begins by defining the system to be analyzed, which is viewed as a composite of subsystems or levels that are integrated to achieve a specific objective. The system is divided into various subsystems or levels, continuing to the lowest possible level, which is referred to as component or element. The system breakdown can either be performed by function (i.e., according to what the system elements "do"), by location (i.e., according to where the system elements "are"), or a combination of both (i.e., functional breakdown by location or vice versa). In a printed circuit board, for example, a location breakdown would include the package, plated through-hole, metallization, and the board itself.

20.5.2 Potential Failure Modes

For all the elements that have been identified, all possible failure modes are listed. For example, in a solder joint, a potential failure mode is open or intermittent change in resistance, which can hamper its functioning as an interconnect. In cases where information on possible failure modes is not available, potential failure modes may be identified using numerical stress analysis, accelerated tests to failure [e.g., highly accelerated life testing (HALT)], past experience, and engineering judgment [14]. A potential failure

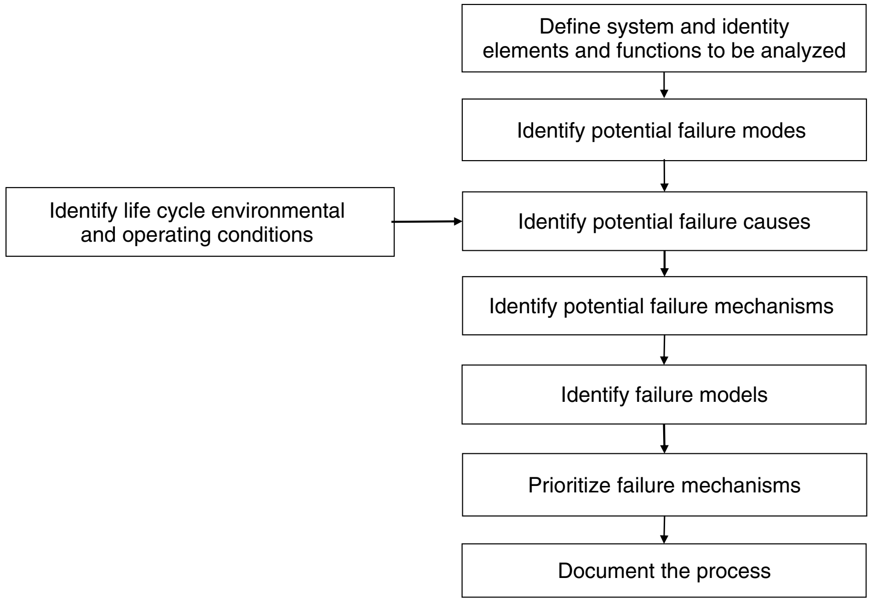


FIGURE 20.4 FMMEA methodology.

mode at one level may be the cause of a potential failure mode in a higher-level system or subsystem, or be the effect of one in a lower level component.

20.5.3 Potential Failure Causes

A failure cause is a circumstance or contributing factor that leads to a failure mode [14]. For each failure mode, all possible ways a failure can result are listed. The failure causes can include environmental and operational conditions. In an automotive underhood environment, for example, circumstances that may lead to solder joint failure modes such as open and intermittent change in resistance can be temperature cycling, random vibration, and shock impact. Knowledge of the potential failure causes can help identify the failure mechanisms that drive the failure modes for a given element.

20.5.4 Potential Failure Mechanisms

Failure mechanisms are the processes by which specific combination of physical, electrical, chemical, and mechanical stresses induce failure [17]. Failure mechanisms are determined based on the combination of potential failure mode and cause of failure [18] and selection of appropriate available mechanisms corresponding to the failure mode and cause. Studies on electronic material failure mechanisms and the application of physics-based damage models to the design of reliable electronic products comprising all relevant wear-out and overstress failures in electronics are available in literature [19,20].

Failure mechanisms thus identified are categorized as either overstress or wear-out mechanisms. Overstress failures involve a failure that arises as a result of a single load (stress) condition. Wear-out failure on the other hand involves a failure that arises as a result of cumulative load (stress) conditions [16]. For example, in the case of solder joint, the potential failure mechanisms driving the opens and shorts caused by temperature, vibration, and shock impact are fatigue and overstress shock. Further analyses of the failure mechanisms depend on the type of mechanism.

20.5.5 Failure Models

Failure models use appropriate stress and damage analysis methods to evaluate susceptibility of failure. Failure susceptibility is evaluated by assessing the TTF or likelihood of a failure for a given geometry, material construction, environmental, and operational condition. For example, in case of solder joint fatigue, Dasgupta [21] and Coffin-Manson [22] failure models are used for stress and damage analysis for temperature cycling.

Failure models of overstress mechanisms use stress analysis to estimate the likelihood of a failure based on a single exposure to a defined stress condition. The simplest formulation for an overstress model is the comparison of an induced stress vs. the strength of the material that must sustain that stress. Wear-out mechanisms are analyzed using both stress and damage analysis to calculate the time required to induce failure based on a defined stress condition. In the case of wear-out failures, damage is accumulated over a period until the item is no longer able to withstand the applied load. Therefore, an appropriate method for combining multiple conditions must be determined for assessing the TTF. Sometimes, the damage due to the individual loading conditions may be analyzed separately, and the failure assessment results may be combined in a cumulative manner.

Failure models may be limited by the availability and accuracy of models for quantifying the TTF of the system. It may also be limited by the ability to combine the results of multiple failure models for a single failure site and the ability to combine results of the same model for multiple stress conditions [14]. If no failure models are available, the appropriate parameter(s) to monitor can be selected based on an empirical model developed from prior field failure data or models derived from accelerated testing (Table 20.1).

20.5.6 Failure Mechanism Prioritization

In the life cycle of a product, several failure mechanisms may be activated by different environmental and operational parameters acting at various stress levels. But, in general, only a few operational and environmental parameters and failure mechanisms, are responsible for a majority of the failures. High-priority failure mechanisms determine the operational stresses, and environmental and operational parameters that must be accounted for in the design or be controlled. Prioritization of the failure mechanisms provides an opportunity for effective utilization of resources.

TABLE 20.1 Examples of Failure Mechanisms, Relevant Loads, and Models for Electronics [1]

Failure Mechanism	Failure Sites	Relevant Loads	Sample Model
Fatigue	Die attach, wirebond/TAB, solder leads, bond pads, traces, vias/PTHs, interfaces	ΔT, T_{mean}, $\mathrm{d}T/\mathrm{d}t$, dwell time, ΔH, ΔV	Nonlinear Power Law (Coffin-Manson)
Corrosion	Metallizations	M, ΔV, T	Eyring (Howard)
Electromigration	Metallizations	T, J	Eyring (Black)
Conductive filament formation	Between metallizations	M, ΔV	Power Law (Rudra)
Stress driven diffusion voiding	Metal traces	s, T	Eyring (Okabayashi)
Time dependent dielectric breakdown	Dielectric layers	V, T	Arrhenius (Fowler-Nordheim)

Δ, Cyclic range; V, voltage; T, temperature; s, stress; Λ, gradient; M, moisture; J, current density; H, humidity; PTH, plated through-hole.

Initial prioritization of all potential failure mechanisms is based on environmental and operating conditions. If the stress levels generated by certain operational and environmental conditions are nonexistent or negligible, the failure mechanisms that are exclusively dependent on those environmental and operating conditions are assigned a "low" risk level and are eliminated from further consideration. For all other failure mechanisms, the failure susceptibility is evaluated by conducting a stress analysis to determine when a failure is expected to be precipitated under the given environmental and operating conditions. To provide a qualitative measure of the effect of a failure, each failure mechanism is assigned a severity rating based on the failure site and failure mode by which it manifests. The same mechanism can have different severity level based on the failure site and mode. The final step prioritizes the failure mechanisms into three risk levels: high, medium, and low.

20.5.7 Documentation

The FMMEA process augments the knowledge base of an organization every time it is employed in an organization. Failure mechanisms specific to products and technologies used are identified, failure models are identified or developed, and risk mitigating design changes are made. Documented history and lessons learned provide a framework for FMMEA of future products or future product versions reducing the time to market and development cost.

20.6 Design Techniques

Once the parts, materials, processes, and stress conditions are identified, the objective is to design a product using parts and materials that have been sufficiently characterized in terms of performance over time when subjected to the manufacturing and application profile conditions. Only through a methodical design approach using PoF analysis, testing, and root-cause analysis can a reliable and cost-effective product be designed.

Design guidelines based on PoF models can also be used to develop tests, screens, and derating* factors. Tests based on PoF models can be designed to measure specific quantities, to detect the presence of unexpected flaws, and to detect manufacturing or maintenance problems. Screens can be designed to precipitate failures in the weak population while not cutting into the design life of the normal population. Derating or safety factors can be determined to lower the stresses for the dominant failure mechanisms. The following sections describe approaches that can be employed in the design phase of electronic equipment to improve reliability.

20.6.1 Protective Architectures

The objective of protective architectures is to enable some form of action, after an initial failure or malfunction, to prevent additional or secondary failures. Protective techniques include the use of fuses and circuit breakers, self-sensing structures, and adjustment structures that correct for parametric shifts.

In designs where safety is an issue, it is generally desirable to incorporate some means for preventing a part, structure, or interconnection from failing or from causing further damage when it fails. Fuses and circuit breakers are examples of elements used in electronic products to sense excessive current drain and disconnect power from the concerned part. Fuses within circuits can also safeguard parts against voltage transients or excessive power dissipation and protect power supplies from shorted parts. Similarly, thermostats can be used to sense critical temperature limiting conditions, and to unpower the product or part of the system until the temperature returns to normal. Self-checking circuitry can also be incorporated to sense abnormal conditions and restore normal conditions, or to activate circuitry that will compensate for the malfunction [9].

*Derating is the practice of subjecting parts to lower electrical or mechanical stresses than they can withstand to increase the life expectancy of the part.

In some instances, it may be desirable to permit partial operation of the product after a part failure, possibly with degraded performance, rather than completely unpower the product. For example, in shutting down a failed circuit whose function is to provide precise trimming adjustment within a deadband of another control product, acceptable performance may be achieved, under emergency conditions, with the deadband control product alone [9].

Sometimes, the physical removal of a part from a product can harm or cause failure in another part by removing load, drive, bias, or control. In such cases, an interlock mechanism may be implemented within the part that is to be removed to shut down or protect other parts. The design should enable function of the product after a failure and should provide the capability for self-sensing and adjustment of parametric drifts to avert failures.

By reducing the number of failures, protective techniques also affect product availability and effectiveness. In addition, protective architectures must be designed by considering the impacts of maintenance, repair, and part replacement. For example, if a fuse protecting a circuit is replaced, the following questions need to be answered: What is the impact when the product is reenergized? What protective architectures are appropriate for post-repair operations? What maintenance guidance must be documented and followed when fail-safe protective architectures have or have not been included?

20.6.2 Stress Margins

A properly designed product should be capable of operating satisfactorily with parts that drift or change with time, and with changes in operating conditions such as temperature, humidity, pressure, and altitude, as long as the parameters remain within their rated tolerances.

Figure 20.5 provides a schematic representation of the hierarchy of product stress limits and margins. The specified tolerance limit is set by the manufacturer to limit the conditions of customer use. The design margin corresponds to the stress value that the product is designed to survive without failure. The operating margin is the expected stress value that may lead to a recoverable failure. The destruct margin is the expected stress value that may lead to permanent (overstress) failure.

To prevent out-of-tolerance failures, a design team must consider the combined effects on stress levels of manufacturing tolerances, including those for parts used in future repair or maintenance operations; expected environmental conditions; and drifts due to aging over the period specified in the reliability requirement. Parts and structures should be designed to operate satisfactorily at the extremes of the parameter ranges and allowable ranges must be included in the procurement or reprocurement specifications.

Statistical analysis and worst-case analysis can be used to assess the effects of part and structural parameter variations. In statistical analysis, a functional relationship is established between the output

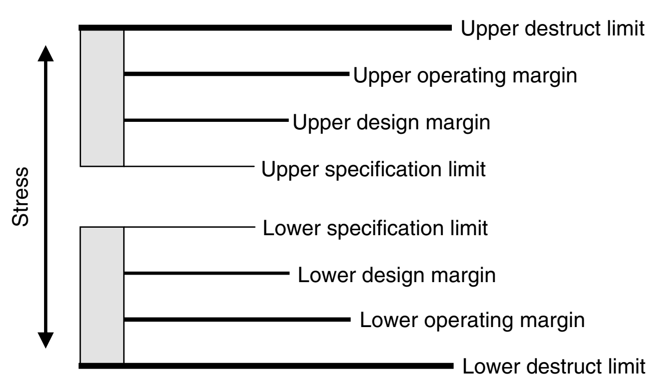


FIGURE 20.5 Stress limits and margins.

characteristics of the structure and the parameters of one or more of its parts. In worst-case analysis, the effect that a part has on product output is evaluated on the basis of end-of-life performance values or out-of-specification replacement parts.

To ensure that the parts used in a system remain within the predetermined margins shown in Figure 20.5, derating can be used. Derating is the practice of limiting thermal, electrical, and mechanical stresses to levels below the manufacturer's specified ratings to improve reliability. Derating can provide added protection from system anomalies unforeseen by the designer (e.g., transient load, electrical surge). For example, manufacturers of electronic hardware often specify limits for supply voltage, output current, power dissipation, junction temperature, and frequency. The equipment design team may decide to select an alternative component or make a design change that ensures that the operational condition for a particular parameter, such as temperature, is always below the rated level. The stress reduction is expected to extend the useful operating life where the failure mechanisms under consideration are of wear-out type. This practice is also expected to provide a safer operating condition by furnishing a margin of safety when the failure mechanisms are of overstress type.

As inherently suggested by the term "derating," the methodology involves a two-step process: a "rated" stress value is first determined from the part manufacturer's data-book, and a reduced value is then assigned. Derating is meant to provide a margin of safety quantified by the difference between maximum allowable applied stress and demonstrated limit of part capability. The part capabilities as given by manufacturer specifications are taken as the demonstrated limits. System design teams often incline to design a product based on conservative stress assumptions due to lack of knowledge of the life cycle conditions at the expense of overall productivity. There are reasons to believe that the part manufacturers already provide safety margin when choosing the operating limits. When these values are derated by the users, effectively a second level of safety margin is added.

In order to be effective, derating must target the appropriate, critical stress parameters to address modeling of the relevant failure mechanisms. Once the failure models for the critical failure mechanisms have been identified using, for example, an FMMEA, the impact of derating on the effective reliability of the part for a given load can be determined. Instead of derating the stress rating values provided by the device manufacturers, the goal should be to determine the safe operating envelope for each part and subsystem and then operate within that envelope.

20.6.3 Redundancy

Redundancy exists when one or more of the parts of a system can fail and the system can still function with the parts that remain operational. A design team often finds that redundancy is the quickest way to improve product reliability if there is insufficient time to explore alternatives, or if the part is already designed. It can be the most cost-effective solution, if the cost of redundancy is economical in comparison with the cost of redesign. It may even be the only solution, if the reliability requirement is beyond the state of the art.

Redundancy may also cause to exceed the limitations on size and weight, or power limitations. When not properly implemented, redundancy can also provide a false sense of security. It is often difficult to assess the performance of redundancy (both active and standby) due to factors including common mode failures, load sharing, and switching and standby failures.

Common mode failures are caused by phenomena that create dependencies between two or more redundant parts, which cause them to fail simultaneously. Common mode failures can be caused by many factors such as common electric connections, shared environmental stresses, and common maintenance problems. In system reliability analysis, common mode failures have the same effect as placing an additional part in series with the parallel redundant configuration.

Load sharing failures occur when the failure of one part increases the stress level of other part(s). This can affect the life of the active part(s). For redundant engines, motors, pumps, and many other systems and devices in active parallel setup, the failure of one part may increase the load on the other parts and decrease their times to failure (or increase their hazard rates).

Several common assumptions are generally made regarding the switching and sensing of standby systems. We assume that switching is in one direction only, that is, the switching devices respond only when directed to switch by the monitor, and that switching devices do not fail if not energized. Regarding standby, the general assumption is that standby nonoperating units cannot fail if not energized. When any of these idealizations are not met, switching and standby failures occur. Monitoring or sensing failure includes both dynamic (failure to switch when active path fails) and static (switching when not required) failures.

Besides these limitations, the design team may find that the disadvantages of redundancy may outweigh the benefits when the implementation of redundancy:

- Prove too expensive with costly redundant sensors and switching devices
- Exceed the limitations on size and weight
- Exceed power limitations, particularly in active redundancy
- Require sensing and switching circuitry so complex as to offset the reliability advantage of redundancy

20.6.4 Integrated Diagnostics and Prognostics

Design guidelines and techniques should involve strategies to assess the reliability of the product in its life cycle environment. A product's health is the extent of deviation or degradation from its expected normal (in terms of physical and performance) operating condition [9]. Knowledge of a product's health can be used for the detection and isolation of faults or failures (diagnostics) and for the prediction of an impending failure based on current conditions (prognostics). Thus, by determining the advent of failure based on actual life cycle conditions, procedures can be developed to mitigate and manage potential failures and maintain the product.

Diagnostics and prognostics can be integrated into a product by (a) installing built-in fuses and canary structures that will fail faster than the actual product when subjected to life cycle conditions [23]; (b) sensing parameters that are precursors to failure, such as defects or performance degradation [24], (c) sensing the life cycle environmental and operational loads that influence the system's health, and processing the measured data to estimate the life consumed [24–26]. The life cycle data measured by such integrated monitors can be extremely useful in future product design and end-of-life decisions [27].

An example of integrated prognostics in electronic products is the self-monitoring analysis and reporting technology currently employed for hard disk drives (HDD) in certain computing equipment [28]. HDD operating parameters, including flying height of the head, error counts; variations in spin time; temperature; and data transfer rates, are monitored to provide advance warning of failures. This is achieved through an interface between the computer's startup program (BIOS) and the HDD.

20.7 Qualification and Accelerated Testing

Qualification includes all activities that ensure that the nominal design and manufacturing specifications will meet the desired reliability targets. Qualification validates the ability of the nominal design and manufacturing specifications of the product to meet the customer's expectations, and assesses the probability of survival of the product over its complete life cycle. The purpose of qualification is to define the acceptable range of variability for all critical product parameters affected by design and manufacturing, such as geometric dimensions, material properties, and operating environmental limits. Product attributes that are outside the acceptable ranges are termed defects, since they have the potential to compromise product reliability [29].

Qualification tests should be performed only during initial product development and immediately after any design or manufacturing changes in an existing product. A well-designed qualification procedure provides economic savings and quick turnaround during development of new products or products subject to manufacturing and process changes.

Investigating failure mechanisms and assessing the reliability of products where longevity is required may be a challenge, since a very long test period under actual operating conditions is necessary to obtain sufficient data to determine actual failure characteristics. The results from FMMEA should be used to guide this process. One approach to the problem of obtaining meaningful qualification data for high-reliability devices in shorter time periods is using methods such as virtual qualification and accelerated testing to achieve test-time compression.

20.7.1 Virtual Qualification

Virtual qualification has the potential to significantly accelerate the qualification process of a part for its life cycle environment. Virtual qualification is based on computer-aided simulation and can assist in identifying and ranking the dominant failure mechanisms associated with the part under life cycle loads, determining the acceleration factor for a given set of accelerated test parameters, and determining the TTF corresponding to the identified failure mechanisms [30].

Each failure model comprises a stress analysis model and a damage assessment model. The output is a ranking of different failure mechanisms, based on the TTF. The stress model captures the product architecture, while the damage model depends on a material's response to the applied stress. Virtual qualification can be used to optimize the product design in such a way that the minimum TTF of any part of the product is greater than its desired life. Although the data obtained from virtual qualification cannot fully replace those obtained from physical tests, they can increase the efficiency of physical tests by indicating the potential failure modes and mechanisms that can be expected.

Ideally, a virtual qualification process will involve identification of quality suppliers, quality parts, PoF qualification, and a risk assessment and mitigation program. The process allows qualification to be readily incorporated into the design phase of product development since it allows design, manufacturing, and testing to be conducted promptly and cost-effectively. It also allows consumers to qualify off-the-shelf components for use in specific environments without extensive physical tests. Since virtual qualification reduces emphasis on examining a physical sample, it is imperative that the manufacturing technology and quality assurance capability of the manufacturer be taken into account. If the data on which the virtual qualification is performed are inaccurate or unreliable, all results are suspect. In addition, if a reduced quantity of physical tests is performed in the interest of simply verifying virtual results, the operator needs to be confident that the group of parts selected is sufficient to represent the product. The effect of manufacturing variabilities can be parametrically assessed by simulation as part of the virtual qualification process. Further, it should be remembered that the accuracy of the results using virtual qualification depends on the accuracy of the inputs to the process—that is, the system geometry and material properties, the life cycle loads, the failure models used (e.g., constants in the failure model), the analysis domain, and numerics (e.g., spatial and temporal discretization) [30,31]. Hence, to obtain a reliable prediction, the variability in the inputs should be specified using distribution functions and the validity of the failure models should be tested by conducting accelerated tests.

20.7.2 Accelerated Testing

Accelerated testing is based on the concept that a product will exhibit the same failure mechanism and mode in a short time under high-stress conditions, as it would exhibit in a longer time under actual life cycle stress conditions. The purpose is to decrease the total time and cost required to obtain reliability information for the product under study. It is generally possible to quantitatively extrapolate from the accelerated environment to the usage environment with some reasonable degree of assurance.

Accelerated tests can be divided into two categories: qualitative tests and quantitative tests. Qualitative tests in the form of overstressing the products to obtain failure may be the oldest form of reliability testing. Such tests typically target a single load condition such as shock, temperature extremes, and electrical overstress. Qualitative tests generally only yield failure mode information and do not reveal the failure mechanism or TTF. Quantitative tests target wear-out failure mechanisms, in which failures occur as a result of cumulative load conditions. TTF is the major outcome of quantitative accelerated tests.

The easiest and most common form of accelerated life testing is continuous-use acceleration. For example, a washing machine is used for 10 h per week in average. If it was continuously operated, the acceleration factor would be 16.8. However, this method is not applicable to high-usage products. Under such circumstances, accelerated testing is conducted to measure the performance of the test product at loads or stresses that are more severe than would normally be encountered in order to enhance the damage accumulation rate within a reduced time period. The goal of such testing is to accelerate time-dependent failure mechanisms and the damage accumulation rate to reduce the TTF. On the basis of the data from those accelerated tests, life in normal use conditions can be extrapolated.

Accelerated testing begins by identifying all the possible overstress and wear-out failure mechanisms. The load parameter that directly causes the time-dependent failure is selected as the acceleration parameter and is commonly called the accelerated load. Common accelerated loads include thermal loads such as temperature, temperature cycling, and rates of temperature change; chemical loads, such as humidity, corrosives, acid, and salt; electrical loads, such as voltage or power; and mechanical loads, such as vibration, mechanical load cycles, strain cycles, and shock/impulses. The accelerated environment may include a combination of these loads. Interpretation of the results for combined loads requires a quantitative understanding of their relative interactions and the contribution of each load to the overall damage.

Failure due to a particular mechanism can be induced by several acceleration parameters. For example, corrosion can be accelerated by both temperature and humidity, and creep can be accelerated by both mechanical stress and temperature [32]. Furthermore, a single accelerated stress can induce failure by several wear-out mechanisms simultaneously. For example, temperature can accelerate wear-out damage accumulation not only by electromigration, but also by corrosion and creep. Failure mechanisms that dominate under usual operating conditions may lose their dominance as the stress is elevated. Conversely, failure mechanisms that are dormant under normal use conditions may contribute to device failure under accelerated conditions. Thus, accelerated tests require careful planning if they are to represent the actual usage environments and operating conditions without introducing extraneous failure mechanisms or nonrepresentative physical or material behavior. The degree of stress acceleration is usually controlled by an acceleration factor defined as the ratio of the life of the product under normal use conditions to that under the accelerated condition. The acceleration factor should be tailored to the hardware in question and can be estimated from an acceleration transform (i.e., a functional relationship between the accelerated stress and the life cycle stress) in terms of all the hardware parameters.

Once the failure mechanisms are identified, it is necessary to select the appropriate acceleration load; to determine the test procedures and the stress levels; to determine the test method, such as constant stress acceleration or step-stress acceleration; to perform the tests; and to interpret the test data, which includes extrapolating the accelerated test results to normal operating conditions. The test results provide failure information for improving the hardware through design and/or process changes.

20.8 Monitor and Control Manufacturing

Manufacturing and assembly processes can significantly impact the quality and reliability of hardware. Improper assembly and manufacturing techniques can introduce defects, flaws, and residual stresses, which act as potential failure sites or stress raisers later in the life of the product. The effect of manufacturing variability on TTF is depicted in Figure 20.6. A shift in the mean or increase in the standard deviation of key geometric parameters during manufacturing can result in early failure due to a decrease in strength of the product. If these defects and stresses can be identified, the design analyst can proactively account for them during the design and development phase.

Auditing the merits of the manufacturing process involves two crucial steps. First, qualification procedures are required, as in design qualification, to ensure that manufacturing specifications do not compromise the long-term reliability of the hardware. Second, lot-to-lot screening is required to ensure that the

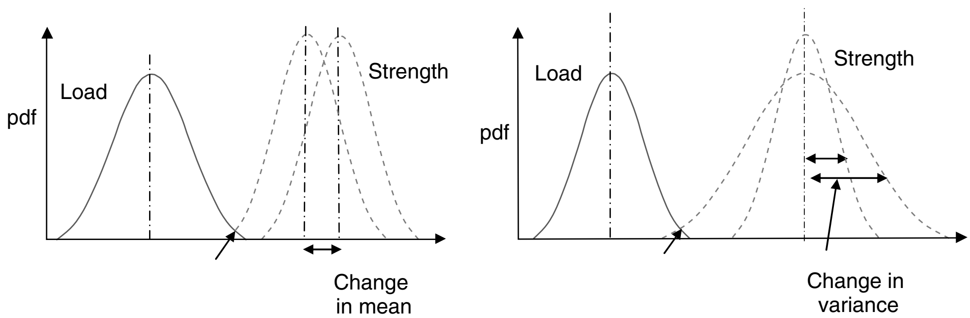


FIGURE 20.6 Influence of quality on failure probability.

variability of all manufacturing-related parameters is within specified tolerances [29,33]. In other words, screening ensures the quality of the product by precipitating latent defects before they reach the field.

20.8.1 Manufacturability

The control and rectification of manufacturing defects has typically been the concern of production and process-control engineers, but not of the design team. In the spirit and context of concurrent product development, however, hardware design teams must understand material limits, available processes, and manufacturing process capabilities to select materials and construct architectures that promote productibility and reduce the occurrence of defects, thus increasing yield and quality. Therefore, no specification is complete without a clear discussion of manufacturing defects and acceptability limits. The reliability engineer must have clear definitions of the threshold for acceptable quality and of what constitutes nonconformance. Nonconformance that compromises hardware performance and reliability is considered a defect. Failure mechanism models provide a convenient vehicle for developing such criteria. It is important for the reliability analyst to understand which deviations from specifications can compromise performance or reliability, and which deviations are benign and can be accepted.

A defect is any outcome of a process (manufacturing or assembly) that impairs or has the potential to impair the functionality of the product at any time. The defect may arise during a single process or may be the result of a sequence of processes. The yield of a process is the fraction of products that are acceptable for use in a subsequent manufacturing sequence or product life cycle. The cumulative yield of the process is approximately determined by multiplying the individual yields of each of the individual process steps. The source of defects is not always apparent, because defects resulting from a process can go undetected until the product reaches some downstream point in the process sequence, especially if screening is not employed.

It is often possible to simplify the manufacturing and assembly processes to reduce the probability of workmanship defects. As processes become more sophisticated, however, process monitoring and control are necessary to ensure a defect-free product. The bounds that specify whether the process is within tolerance limits, often referred to as the process window, are defined in terms of the independent variables to be controlled within the process and the effects of the process on the product or the dependent product variables. The goal is to understand the effect of each process variable on each product parameter to formulate control limits for the process—that is, the points on the variable scale where the defect rate begins to have a potential for causing failure. In defining the process window, the upper and lower limits of each process variable beyond which it will produce defects must be determined. Manufacturing processes must be contained in the process window by defect testing, analysis of the causes of defects, and elimination of defects by process control, such as by closed-loop corrective action systems. The establishment of an effective feedback path to report process-related defect data is critical. Once this is established and the process window is determined, the process window itself becomes a feedback system for the process operator.

Several process parameters may interact to produce a different defect than would have resulted from an individual parameter acting independently. This complex case may require that the interaction of various process parameters be evaluated in a matrix of experiments. In some cases, a defect cannot be detected until late in the process sequence. Thus, a defect can cause rejection, rework, or failure of the product after considerable value has been added to it. These cost items due to defects can reduce a return on investment by adding to hidden factory costs. All critical processes require special attention for defect elimination by process control.

20.8.2 Process Qualification

Similar to design qualification, process qualification should be conducted at the prototype development phase. The objective is to ensure that the nominal manufacturing specifications and tolerances produce acceptable reliability in the product. The process needs requalification when process parameters, materials, manufacturing specifications, or human factors change.

Process qualification tests can be the same set of accelerated wear-out tests used in design qualification. As in design qualification, overstress tests may be used to qualify a product for anticipated field overstress loads. Overstress tests may also be exploited to ensure that manufacturing processes do not degrade the intrinsic material strength of the hardware beyond a specified limit. However, such tests should supplement, not replace, the accelerated wear-out test program, unless explicit physics-based correlations are available between overstress test results and wear-out field-failure data.

20.8.3 Process Verification Testing

Process verification testing is often called screening. Screening involves 100% auditing of all manufactured products to detect or precipitate defects. The aim of this step is to preempt potential quality problems before they reach the field. Thus, screening aids in reducing warranty returns and increases customer goodwill. In principle, screening should not be required for a well-controlled process. However, screening is often used as a safety net.

Some products exhibit a multimodal probability density function for failures, as depicted in Figure 20.7, with peaks during the early period of their service life due to the use of faulty materials, poorly controlled manufacturing and assembly technologies, or mishandling. This type of early-life failure is often called infant mortality. Properly applied screening techniques can successfully detect or precipitate these failures, eliminating or reducing their occurrence in field use. Screening should only be considered for use during the early stages of production, if at all, and only when products are expected to exhibit infant mortality field failures. Screening will be ineffective and costly if there is only one main peak in the failure probability

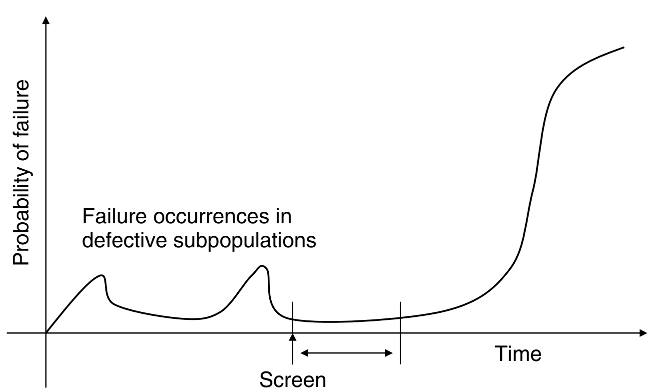


FIGURE 20.7 Candidate for screening due to wear-out failure.

density function. Further, failures arising due to unanticipated events such as acts of God (lightning, earthquakes) may be impossible to screen cost-effectively.

Since screening is performed on a 100% basis, it is important to develop screens that do not harm good components. The best screens, therefore, are nondestructive evaluation techniques, such as microscopic visual exams, X-rays, acoustic scans, nuclear magnetic resonance, electronic paramagnetic resonance, and so on. Stress screening involves the application of stresses, possibly above the rated operational limits. If stress screens are unavoidable, overstress tests are preferred over accelerated wear-out tests, since the latter are more likely to consume some useful life of good components. If damage to good components is unavoidable during stress screening, then quantitative estimates of the screening damage, based on failure mechanism models, must be developed to allow the design team to account for this loss of usable life. The appropriate stress levels for screening must be tailored to the specific hardware. As in qualification testing, quantitative models of failure mechanisms can aid in determining screen parameters.

A stress screen need not necessarily simulate the field environment or even utilize the same failure mechanism as the one likely to be triggered by this defect in field conditions. Instead, a screen should exploit the most convenient and effective failure mechanism to stimulate the defects that can show up in the field as infant mortality. Obviously, this requires an awareness of the possible defects that may occur in the hardware and extensive familiarity with the associated failure mechanisms.

Unlike qualification testing, the effectiveness of screens is maximized when screening is conducted immediately after the operation believed to be responsible for introducing the defect. Qualification testing is preferably conducted on the finished product or as close to the final operation as possible; on the other hand, screening only at the final stage, when all operations have been completed, is less effective, since failure analysis, defect diagnostics, and troubleshooting are difficult and impair corrective actions. Further, if a defect is introduced early in the manufacturing process, subsequent value added through new materials and processes is wasted, which additionally burdens operating costs and reduces productivity.

Admittedly, there are also several disadvantages to such an approach. The cost of screening at every manufacturing station may be prohibitive, especially for small batch jobs. Further, components will experience repeated screening loads as they pass through several manufacturing steps, which increases the risk of accumulating wear-out damage in good components due to screening stresses. To arrive at a screening matrix that addresses as many defects and failure mechanisms as feasible with each screen test, an optimum situation must be sought through analysis of cost-effectiveness, risk, and the criticality of the defects. All defects must be traced back to the root cause of the variability.

Any commitment to stress screening must include the necessary funding and staff to determine the root cause and appropriate corrective actions for all failed units. The type of stress screening chosen should be derived from the design, manufacturing, and quality teams. Although a stress screen may be necessary during the early stages of production, stress screening carries substantial penalties in capital, operating expense, and cycle time, and its benefits diminish as a product approaches maturity. If almost all of the products fail in a properly designed screen test, the design is probably faulty. If many products fail, a revision of the manufacturing process is required. If the number of failures in a screen is small, the processes are likely to be within tolerances and the observed faults may be beyond the resources of the design and production process.

20.9 Closed Loop Monitoring

Product reliability needs to be ensured using a closed-loop process that provides feedback to design and manufacturing in each stage of the product life cycle, including after the product is shipped and is used in its application environment. Data obtained from periodic maintenance, inspection/testing, or health (condition) and usage monitoring methods can be used to perform timely maintenance for sustaining the product and preventing catastrophic failures. Figure 20.8 depicts the closed-loop process for managing the reliability of a product over the complete life cycle.

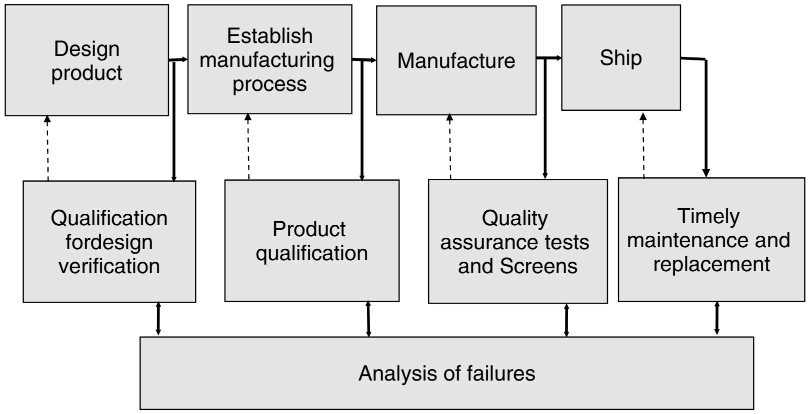


FIGURE 20.8 Reliability management using a closed-loop process.

The objective of closed-loop monitoring is to analyze the failures occurring in testing and field conditions to identify the root cause of failure. The root cause is the most basic casual factor or factors that, if corrected or removed, will prevent recurrence of the situation. The purpose of determining the root cause(s) is to fix the problem at its most basic source so it does not occur again, even in other products, as opposed to merely fixing a failure symptom.

Root-cause analysis is a methodology designed to help (a) describe what happened during a particular occurrence; (b) determine how it happened; and (c) understand why it happened. Only when we are able to determine why an event or failure occurred, will we are able to determine corrective measures over time. Root-cause analysis is different from troubleshooting, which is generally employed to eliminate a symptom in a given product, as opposed to finding a solution to the root cause to prevent this and other products from failing.

Correctly identified root cause during design and manufacturing, followed by appropriate actions to fix the design and processes, results in fewer field returns, major cost savings, and customer goodwill. Root-cause analysis of field failures can be more challenging if the conditions during and prior to failure are unknown. The lessons learned from each failure analysis need to be documented and appropriate actions need to be taken to update the design, manufacturing process, and/or maintenance actions or schedules.

20.10 Summary

Reliability is not a matter of chance or good fortune; rather, it is a rational consequence of conscious, systematic, rigorous efforts to ensure reliability through the entire life cycle of the product. High product reliability can only be assured through robust product designs, capable processes that are known to be within tolerances, and qualified components and materials from vendors whose processes are also capable and within tolerances. Quantitative understanding and modeling of all relevant failure mechanisms can guide the design process, manufacturing processes, and the definition of test specifications and tolerances.

The PoF approach is not only a tool to provide better and more effective designs, but it also helps develop cost-effective approaches for improving the entire approach to building electronic products. Proactive actions can be implemented for defining more realistic performance requirements and environmental conditions, identifying and characterizing key material properties, developing new product architectures and technologies, developing more realistic and effective accelerated stress tests to audit reliability and quality, enhancing manufacturing-for-reliability through mechanistic process modeling, and characterization allowing proactive process optimization, increasing first-pass yields, and reducing hidden factory costs associated with inspection, rework, and scrap.

When utilized early in the concept stage of a product's development, reliability serves as an aid to determining feasibility and risk. In the design stage of product development, reliability analysis involves methods to enhance performance over time through the selection of materials, design of structures, design tolerances, manufacturing processes and tolerances, assembly techniques, shipping and handling methods, and maintenance and maintainability guidelines. Engineering concepts such as strength, fatigue, fracture, tolerances, corrosion, and aging play a role in these design analyses. The use of PoF concepts coupled with mechanistic and probabilistic techniques are often required to understand the potential problems and trade-offs, and to take corrective actions. The use of factors of safety and worst-case studies as part of the analysis is useful in determining stress screening and burn-in procedures, reliability growth, maintenance modifications, field-testing procedures, and various logistics requirements.

References

1. Lall, P., Pecht, M., and Hakim, E. *Influence of Temperature on Microelectronics and System Reliability*, CRC Press, New York, 1997.
2. Pecht, M. "Why traditional reliability prediction models do not work—is there an alternative," *Electronics Cooling*, Vol. 2, No. 1, pp. 10–12, 1996.
3. Das, D. "Use of thermal analysis information in avionics equipment development," *Electronics Cooling*, Vol. 5, No. 3, pp. 28–34, 1999.
4. Osterman, M. "We still have a headache with Arrhenius," *Electronics Cooling*, Vol. 7, No. 1, pp. 53–54, 2001.
5. Dasgupta, A. "The physics-of-failure approach at the University of Maryland for the development of reliable electronics," *Proceedings of the Third International Conference on Thermal and Mechanical Simulation in (Micro) Electronics* (*EuroSimE*), pp. 10–17, 2002.
6. Jackson, M., Mathur A., Pecht M., and Kendall R. "Part manufacturer assessment process," *Quality and Reliability Engineering International*, Vol. 15, pp. 457–468, 1999.
7. Jackson, M., Sandborn, P., Pecht, M., Hemens-Davis, C., and Audette, P., "A risk-informed methodology for parts selection and management," *Quality and Reliability Engineering International*, Vol. 15, pp. 261–271, 1999.
8. IEEE Standard 1332, *IEEE Standard Reliability Program for the Development and Production of Electronic Systems and Equipment*, 1998.
9. Vichare, N., Rodgers, P., Eveloy, V., and Pecht, M. "*In situ* temperature measurement of a notebook computer—a case study in health and usage monitoring of electronics," *IEEE Transactions on Device and Materials Reliability*, Vol. 4, No. 4, pp. 658–663, 2004.
10. Lewis E.E. *Introduction to Reliability Engineering*, John Wiley & Sons, New York, 1996.
11. IEC Standard 60134, *Rating Systems for Electronic Tubes and Valves and Analogous Semiconductor Devices* (Last reviewed in July 1994 by the IEC Technical Committee 39 on Semiconductors), 1961.
12. Humphrey, D., Condra, L., Pendse, N., Das, D., Wilkinson, C., and Pecht, M. "An avionics guide to uprating of electronic parts," *IEEE Transactions on Components and Packaging Technologies*, Vol. 23, No. 3, pp. 595–599, 2000.
13. SAE Standard SAE J1739, *Potential Failure Mode and Effects Analysis in Design (Design FMEA) and Potential Failure Mode and Effects Analysis in Manufacturing and Assembly Processes (Process FMEA) and Effects Analysis for Machinery (Machinery FMEA)*, August 2002.
14. *Guidelines for Failure Mode and Effects Analysis for Automotive, Aerospace and General Manufacturing Industries*, Dyadem Press, Ontario, Canada, 2003.
15. Bowles, J.B. *Fundamentals of Failure Modes and Effects Analysis*, Tutorial Notes Annual Reliability and Maintainability Symposium, 2003.
16. Ganesan, S., Eveloy, V., Das, D., and Pecht, M.G., "Identification and utilization of failure mechanisms to enhance FMEA and FMECA," *Proceedings of the IEEE Workshop on Accelerated Stress Testing and Reliability* (*ASTR*), Austin, TX, October 3–5, 2005.
17. Hu, J., Barker, D., Dasgupta, A., and Arora, A., "Role of failure-mechanism identification in accelerated testing," *Journal of the IES*, Vol. 36, No. 4, pp. 39–45, 1993.
18. JEDEC, JEDEC Publication JEP 148 "Reliability qualification of semiconductor devices based on physics of failure risk and opportunity assessment," April 2004.

19. Dasgupta, A. and Pecht, M. "Material failure mechanisms and damage models," *IEEE Transactions on Reliability*, Vol. 40, No. 5, pp. 531–536, 1991.
20. JEDEC, JEDEC Publication JEP 122-B "Failure mechanisms and models for semiconductor devices," August 2003.
21. Dasgupta, A., Oyan, C., Barker, D., and Pecht, M., "Solder creep-fatigue analysis by an energy-partitioning approach," *Transactions of the ASME, Journal of Electronic Packaging*, Vol. 114, No. 2, pp. 152–160, 1992.
22. Foucher, B., Boullie, J., Meslet, B., and Das, D. "A review of reliability predictions methods for electronic devices," *Microelectronics Reliability*, Vol. 42, No. 8, pp. 1155–1162, 2002.
23. Mishra, S. and Pecht, M. "*In-situ* sensors for product reliability monitoring," *Proceedings of SPIE*, Vol. 4755, pp. 10–19, 2002.
24. Pecht, M., Dube, M., Natishan, M. and Knowles, I. "An evaluation of built-in test," *IEEE Transactions on Aerospace and Electronic Systems*, Vol. 37, No. 1, pp. 266–272, 2001.
25. Mishra, S., Pecht, M., Smith, T., McNee, I., and Harris, R. "Remaining life prediction of electronic products using life consumption monitoring approach," *Proceedings of the European Microelectronics Packaging and Interconnection Symposium*, Cracow, Poland, 16–18 June 2002, pp. 136–142.
26. Ramakrishnan, A. and Pecht, M. "A life consumption monitoring methodology for electronic systems," *IEEE Transactions on Components and Packaging Technologies*, Vol. 26, No. 3, pp. 625–634, 2003.
27. Vichare, N., Rodgers, P., Azarian, M.H., and Pecht, M. "Application of health monitoring to product take-back decisions," *Joint International Congress and Exhibition—Electronics Goes Green 2004+*, Berlin, Germany, 6–8 September 2004.
28. Self-Monitoring Analysis and Reporting Technology (SMART), PC Guide, http://www.pcguide.com/ref/hdd/perf/qual/featuresSMART-c.html, last accessed on 22 August 2004.
29. Pecht, M., Dasgupta, A., Evans, J.W., and Evans, J.Y. *Quality Conformance and Qualification of Microelectronic Packages and Interconnects*, John Wiley & Sons, New York, 1994.
30. Mathew, S., Das, D., Osterman, M., Pecht, M. and Ferebee, R. "Prognostic assessment of aluminum support structure on a printed circuit board," *ASME Journal of Electronic Packaging*, Vol. 128, No. 4, pp. 339–345, 2006.
31. Rodgers, P., Eveloy, V., and Hashmi, M. "An investigation into the potential of low-reynolds number eddy viscosity turbulent flow models to predict electronic component operational temperature," *Transactions of the ASME, Journal of Electronic Packaging*, Vol. 127, No. 1, pp. 67–75, 2005.
32. JEDEC, JEP 122B, *Failure Mechanisms and Models for Semiconductor Devices*, JEDEC Standard, August 2003.
33. Upadhyayula, K. and Dasgupta, A. "Guidelines for physics-of-failure based accelerated stress testing," *Proceedings of Annual Reliability Maintainability Symposium*, New York, NY, pp. 345–357, 1998.

21 Safety and Environmental Issues

John M. Osepchuk
Full Spectrum Consulting

Ronald C. Petersen
R C Petersen Associates LLC

21.1 Characteristics of Biological Tissue and RF Absorption Properties

Modern IEEE standards for safe exposure to radio frequency (RF) electromagnetic energy cover frequencies up to 300 GHz and down to at least 3 kHz. Although the term "microwaves" usually means frequencies well above 100 or 300 MHz, related bioeffects/hazards exist down to roughly 100 kHz. These are thermal in nature. Below 100 kHz, the dominant effects are electrostimulation in nature. We will confine our attention to thermal effects and refer the reader to an authoritative treatment on the subject of electrostimulation by Reilly [1,2].

In order to understand any bioeffect caused by exposure of a biological body to microwave/RF energy, one needs to have an idea of the distribution of the internal E and B fields generated by the exposure. In turn, one must have some information about the dielectric properties of the tissues in the biological body. Later we will refer to modern computer modeling of the absorption using elaborate anatomically correct models of the body, which is heterogeneous in general with muscle, bone, skin, and so forth. Some simple but important properties of the absorption can be gained by simple models.

The dielectric properties of various tissues have been tabulated in popular references including the *Radiation Dosimetry Handbook*, edited by Durney et al. [3]. In recent years, these data, particularly for bone, have been improved by Gabriel et al. [4]. From these sources, we can tabulate the approximate values for muscle-like tissue as in Table 21.1.

TABLE 21.1 Approximate Dielectric Parameters for Muscle Tissue at Various Frequencies[a]

Frequency (MHz)	Relative Dielectric Constant (ε_r)	Conductivity (σ) (S/m)	Penetration Depth (δ) (cm)
0.1	1850	0.56	213
1.0	411	0.59	70
10	131	0.68	13.2
100	79	0.81	7.7
1000	60	1.33	3.4
10,000	42	13.3	0.27
100,000	8	60	0.03

[a] Muscle tissue, field parallel to tissue fibers.

In Table 21.1, the relative dielectric constant is shown as well as the conductivity. The complex permittivity is given by:

$$\varepsilon = \varepsilon_r \varepsilon_0 + j\frac{\sigma}{\omega} \tag{21.1}$$

where $\varepsilon_0 = 8.86 \times 10^{-12}$ f/m. The penetration depth is that at which a plane wave is attenuated by a factor of e in E-field or 8.69 dB. Its classical derivation is presented in Durney et al. [3]. We see that penetration at low-RFs is considerably more than 10 cm, but above 6 GHz, the penetration depth rapidly decreases to a millimeter or less in the millimeter-wave range of the microwave spectrum. The penetration depth is one basic factor in determining how much energy reaches deep into the body. The other factor is the reflection at the external surface, or at lower frequencies the shunting of the electric field by a conducting body. For a small spherical object, Schwan [5] has shown that at 60 Hz the internal E-field is nearly six orders of magnitude less than the external E-field, even though the theoretical penetration depth is quite large at low frequencies. It has been estimated by Osepchuk [6] that only around the "resonance" frequency of humans (i.e., around 100 MHz), is the internal E-field deep in the body within one order of magnitude of the external field. At very high frequencies, for example, in the millimeter-wave range, the E-field deep in the body is many orders of magnitude below the external field because the penetration depth is only 1 mm or less.

The principles of modern dosimetry have recently been reviewed by Chou et al. [7]. The specific absorption rate (SAR), that is, the mass averaged rate of energy absorption in tissue, is related to E by

$$\mathrm{SAR} = \frac{\sigma |E|^2}{\rho} \quad \mathrm{W/kg} \tag{21.2}$$

where σ is the conductivity of the tissue in Siemens per meter, ρ is the density in kg/m^3, and E is the rms electric field strength in volts per meter. SAR thus is a measure of the electric field at the point under study and it is also a measure of the local heating rate $\mathrm{d}T/\mathrm{d}t$, namely,

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \frac{\mathrm{SAR}}{c} \quad {}^\circ\mathrm{C/s} \tag{21.3}$$

where c is the specific heat capacity of the tissue in J/kg °C. Thus, an SAR of 1 W/kg is associated with a heating rate less than 0.0003°C/s in muscle tissue ($c \cong 3.5$ kJ/kg°C). Clearly, this is a very small heating rate since even without blood or other cooling factors it would take more than 1 h to increase the temperature by 1°C.

The SAR concept is a key concept in planning and analysis of experiments, both *in vivo* and *in vitro*, as well as the formulation of safe exposure limits for humans. Both the local SAR value and the whole-body average are important in these endeavors. There is an extensive literature on the calculation of whole-body

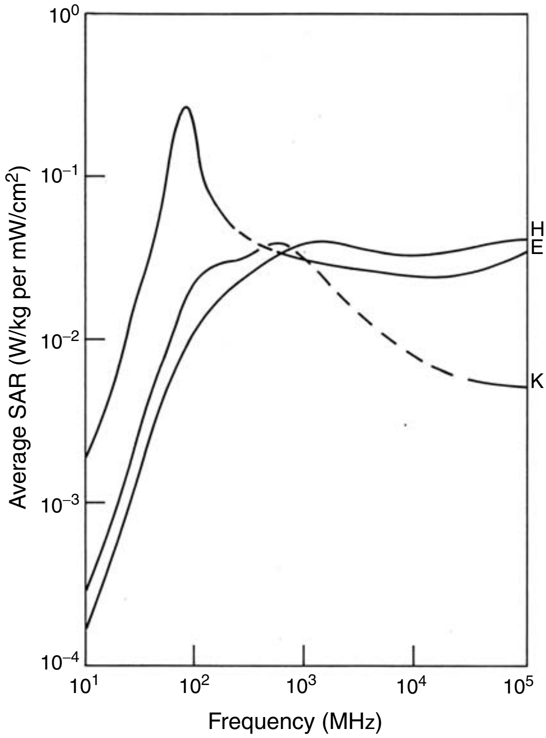


FIGURE 21.1 Calculated whole-body average SAR versus frequency for simple models of the average human for three standard polarizations. The incident power density is 1 mW/cm^2. (From Durney, C. H., et al. *Radiofrequency Radiation Dosimetry Handbook*, Fourth Edition, Report USAFSAM-TR-85-73, USAF School of Aerospace Medicine, Brooks AFB, TX, 1986.)

average SAR for various models of animals, including man, especially those based on ellipsoids, which are summarized in the *Dosimetry Handbook* [3]. In Figure 21.1, we show the calculated SAR for an average human based on such a model, when exposed to a plane wave of power density of 1 mW/cm^2. Shown is the whole-body-averaged SAR versus frequency for three polarizations; *E*—in which the E-field is parallel to the main axis, *H*—in which the H-field is parallel to the main axis, and *k*—in which the direction of propagation is parallel to the main axis of the body. We see that there is a low-Q resonance at about 70–80 MHz for standard human (and at about half that frequency when standing on a conducting ground plane). The peak absorption is highest for *E* polarization and is equal to about 0.2 W/kg per mW/cm^2 incident power density. At high frequencies, the SAR decreases to an asymptotic "quasi-optical" value 5–6 times lower than the SAR peak. At very low frequencies, the SAR varies as f^2, as expected. The peak SAR values for animals are higher. For example, for a mouse at its resonance frequency of about 2 GHz, the peak SAR is somewhat over 1.0 W/kg per mW/cm^2.

The SAR distributions within the body are quite complicated even when resulting from exposure to the simplest plane wave in the far field. Depending on the body position and frequency, it is possible that there are one or more "hot spots" of SAR peaks within the animal body. On the basis of the work of Kritikos and Schwan [8], however, such internal SAR peaks are very unlikely for human but are more probable for small animals.

The rapid development of computer models [9] for electromagnetic fields has been widely applied to calculation of the fields inside the human body under irradiation by RF energy. In many of these studies, detailed models of the human body, some labeled anatomically correct, are employed and various databases on such models are available [10–12]. The nature of the applications includes medical applications, *in-vivo* studies, as well as *in vitro* studies [13,14].

21.2 Bioeffects and Hazards of Microwave/RF Energy

There are estimated to be well above 20,000 papers in the world literature on the bioeffects of microwave/RF energy. Many deal with experiments with animals, that is, *in vivo*, especially at frequencies between 10 MHz and 60 GHz. The early literature before 1980 showed that a major share of these experiments with small animals was done at 2.45 GHz, which, as the microwave-oven frequency, is an inexpensive source of power. During the past decade, however, the research has mainly focused on frequencies used for personal wireless communications services, that is, 800–1000 MHz and 1900–2100 MHz, because of the widespread use of these services. There is also a substantial literature on *in vitro* experiments in which small samples of tissue or cells are exposed in a variety of exposure systems. These occur more after 1980. Epidemiological studies are few and tend to have been done in recent years. For most, exposure assessment is at best questionable and at worst nonexistent. One should not forget that a large amount of human exposure data exist from the past history [15] of diathermy at 27 MHz and at 2.45 GHz. Millions of patients were treated with up to 125 W of power applied to various parts of the body for 15–30 min. In recent years, diathermy has been less popular, perhaps because of the growth of electrophobia directed toward the "radiation" aspect of "microwave" energy. In the past few decades, however, these human exposure data have been tremendously expanded by the astronomical growth in the use of personal wireless devices, for example, "cellular phones" and in the medical area, by the widespread use of MRI (magnetic resonance imaging) in which RF energy at very high frequencies (VHF) and of the order of 100 W is applied to the body. Finally, there are a variety of other medical applications [16] of microwave/RF energy such as hyperthermia for the treatment of cancer.

To better understand experiments in microwave exposure as well as their relation to safety standards, it is useful to refer to the "exposure diagram" of Figure 21.2. In this diagram, with log–log coordinates of power (or power density or SAR) on the ordinate and time on the abscissa, we can draw the threshold for various effects and hazards. For example, to heat a finite sample to a given temperature, the threshold is a constant SAR for long periods of time while for small periods of time, during which no heat is lost from the sample, the threshold curve is a line of constant SA (specific absorption = SAR $\times$ time) which is at 45° from the horizontal in Figure 21.2. The intersection of the two lines, constant SAR and constant SA,

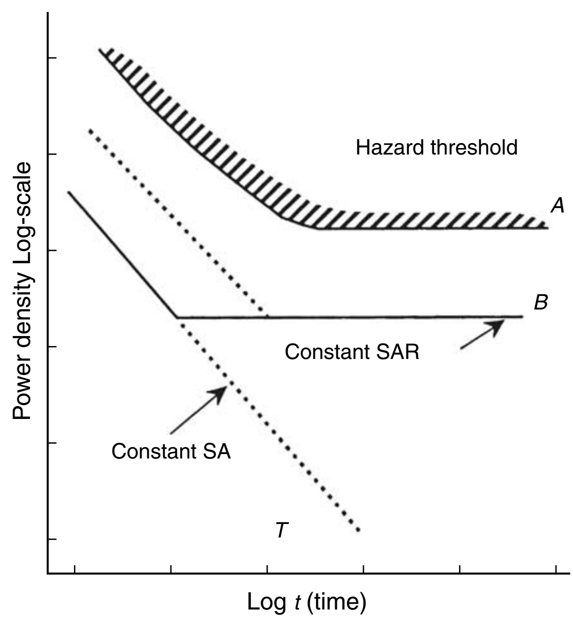


FIGURE 21.2 Thresholds for various effects and hazards expressed as a function of time.

determines the applicable thermal time constant or associated "averaging time" in exposure standards. Similar curves would result for the threshold for burns using the classic data of Moritz and Henriques [17] for threshold temperature for burns, which is around 60°C for 5 s but approaching 45°C for long exposure times, where 45°C is also the threshold temperature for pain sensation in man.

Thus, much early data denoted by such threshold curves described the lethal thresholds of exposure of animals in terms of power density and exposure duration, usually at room temperature. For example, Michaelson et al. [18] found that for a dog, the lethal threshold at 2.8 GHz was about 2–4 h at 165 mW/cm^2. On the other hand, Addington et al. [19] found that at 200 MHz, for a dog the lethal threshold was only 20 min at 220 mW/cm^2. In the Soviet literature [20], the lethal threshold reported for the rat was 40 mW/cm^2 for 90 min at 3 GHz, but at 70 MHz, the threshold was about 1000 mW/cm^2 for 100 min. Although not appreciated at those early times, these results appear quite reasonable in terms of the expected SAR absorption curves and respective resonance frequencies for the various size animals. This physical understanding based on heating was further strengthened when experiments [21] with fruit flies (Drosophilae) showed no effect when exposed at 2.45 GHz with over 6500 mW/cm^2 and 45-min exposure duration. This result is eminently reasonable to the engineer who is well acquainted with the absorption cross section theory that shows absorption decreases rapidly as the square of the animal dimension. (It also explains the mystifying—to the layman—observation that small ants are not perturbed in an operating microwave oven.) In 1971, Samaras et al. [22] demonstrated the expected, but still dramatic, dependence on environmental temperature. At room temperature, the lethal threshold for a rat at 2.45 GHz for 17-min exposure duration was 100 mW/cm^2 but at freezing temperatures below 0°C, that same power density was life preserving for the rat.

In 1979, Tell and Harlen [23] analyzed a host of data in the literature demonstrating thermal effects in animals. They showed how all the animal data seemed to form a coherent picture and extrapolated to humans so as to suggest that, at least for frequencies above 1 GHz, 100 mW/cm^2 was a conservative estimate as the threshold for a 1°C core temperature rise for exposure durations of more than an hour. Their analysis suggested also that the thermal time constant for whole-body heating of man is an hour or more. In the past 15 years, however, actual experiments with humans by Adair et al. [24–26] have shown that whole-body exposures at 100 and 220 MHz and partial-body exposures at 450 MHz and 2.45 GHz for 45 min showed no rise in core temperature at power densities far above the basis of the current safety limits (expressed as 0.4 or 0.08 W/kg). Adair found spatial peak SARs as high as 15 W/kg and found that sweating was often a key mechanism in maintaining thermoregulation. In addition, she concluded that thermal discomfort at the highest exposure levels suggested that such discomfort would motivate one to leave a place of strong RF fields well before thermoregulation ceased to be effective.

Many endpoints of the health of animals have been studied in the thousands of experiments reported in the literature. We will not specifically review all categories but special mention of "cataracts" is justified in view of the myths that were attached to this subject in the 1970s. Cataracts, that is, opacities of the lens that interfere with vision, are probably the most studied effect associated with exposure to microwave energy. The hard science on the subject, for example that of Carpenter, shows that the threshold [27] for cataracts in the rabbit is of the order of 180 mW/cm^2 for an exposure duration of 30 min or more at 2.45 GHz. This result was obtained when the rabbit is restrained or under anesthesia, and only when the energy is applied locally to the eyes with near-field applicators like the diathermy antennas. At X-band, attempts to produce cataracts first resulted in skin burns around the eyes. Attempts to produce cataracts at ultra high frequency (UHF) resulted in the death of the animal before the cataract could be produced. Long-term exposures of the rabbit by Guy et al. [28] at 2.45 GHz and 10 mW/cm^2 showed no ocular damage. Microwave-induced cataracts have not been demonstrated in primates but in the past two decades, there have been some reports [29] of corneal damage from high peak power pulsed fields at moderate average power densities around 10 mW/cm^2. Attempts to replicate these findings have failed [30]. In summary, ocular effects associated with exposure to microwave energy requires a significant temperature increase in the eye, which could only occur at exposure levels far in excess of the limits in contemporary safety standards.

The extensive literature on microwave bioeffects has often been surveyed. Some of the classic papers are reproduced, along with extensive bibliographies and commentaries in a Reprint Volume [31] produced by the IEEE Committee on Man and Radiation (COMAR). Other good reviews of the literature before 1980 include a special issue [32] of the *Proceedings of the IEEE* and an extensive review [33] by the Environmental Protection Agency (EPA). Over the years, Polson and Heynick [34] have produced critical reviews of the literature, under the sponsorship of the U.S. Air Force. An excellent text was authored by Michaelson and Lin [35] and multiauthor books have been edited by Gandhi [36] and Polk and Postow [37]. More recently, the results of extensive literature reviews, which were part of the process underlying the revision of IEEE C95.1 [38], were published [39]. All subject areas including various physiological effects, epidemiology, and *in vitro* studies—for example, on genotoxicity, were covered. The literature reporting research of good quality continues to support contemporary standards and guidelines such as IEEE C95.1-2005 [38] and the International Commission on Non-Ionizing Radiation Protection (ICNIRP) guidelines [40].

Although most confirmed bioeffects are associated with significant temperature rise in experimental animals, there is one exception, that is, the microwave auditory effect [41]. It has been shown that exposure of the head to microwave pulses results in audible clicks above a threshold of roughly 40 μJ/cm^2 energy density incident on the head at 2.45 GHz. This effect is not believed to be hazardous but it has been used in some safety guidelines to set limits for exposures to pulsed fields. Special mention should be made of de Lorge's [42] studies of the disruption of food-motivated learned behavior of animals. Thresholds for this effect, which is believed to be the most sensitive and reproducible known effect, have been the principal basis for most modern safety standards beginning with the series of C95 standards produced under IEEE sponsorship. Disruption occurs reliably at whole-body-averaged SARs between 2–3 and 9 W/kg across frequency and animal species from mice to baboons (see Table 21.2) [38,43].

It needs to be said that this field has been fraught with much confusion and much poor-quality literature. In 1977, Senator Stevenson [44] stated, "I have never gotten into a subject on which there has been so much disagreement and so much confessed lack of knowledge." He was speaking of the field of microwave bioeffects. Later, Guy and Foster [45] and Foster and Pickard [46] wrote critical papers pointing out the prevalence of many papers in the literature that could not be replicated or confirmed. They wondered when does such research that never finds robust confirmed effects cease. An example of such literature is that of the former Soviet Union, and to some extent Germany, which reports frequency-sensitive effects of millimeter-wave radiation at low levels of around 1 mW/cm^2. These reports led to an extensive practice of millimeter-wave therapy for medical purposes in Russia and the Ukraine but neither the research nor the medical practice has been found valid in the West (see the discussion in [46]). More recently, there was a report [47] from Russia of bioeffects at extraordinary low levels of power density $\sim 10^{-19}$ W/cm^2 at a millimeter-wave frequency. We [48] have shown, however, that this extraordinary claim is most probably invalid because of the lack of control of significant energy at the harmonic frequencies. This is only one example of the presence of microwave artifacts that mar many of the papers in the literature.

TABLE 21.2 Comparison of Power Density and SAR Thresholds for Behavioral Disruption in Trained Laboratory Animals

Species and Conditions	225 MHz (CW)	1.3 GHz (Pulsed)	2.45 GHz (CW)	5.8 GHz (Pulsed)
Norwegian rat				
Power density	—	10 mW/cm^2	28 mW/cm^2	20 mW/cm^2
SAR	—	2.5 W/kg	5.0 W/kg	4.9 W/kg
Squirrel monkey				
Power density	—	—	45 mW/cm^2	40 mW/cm^2
SAR	—	—	4.5 W/kg	7.2 W/kg
Rhesus monkey				
Power density	8 mW/cm^2	57 mW/cm^2	67 mW/cm^2	140 mW/cm^2
SAR	3.2 W/kg	4.5 W/kg	4.7 W/kg	8.4 W/kg

Other artifacts include the great nonuniformity of microwave heating of objects, which is often neglected. Thus, often it is reported that the object temperature is some value when in actuality the object has a wide spatial variation in temperature. Unfortunately, these artifacts not only appear in *in vitro* studies but also in *in vivo* studies that have become increasingly prevalent in this field. One recent large study by Repacholi et al. [49] of a large number of transgenic mice has implied a connection of low-level microwave exposure with a cancer (lymphoma). Unfortunately, the experiment was done in a metal enclosure and it is known that exposures in metal cavities, lightly loaded, most probably are chaotic and unpredictable. A subsequent study by Utteridge et al. [50] performed at multiple exposure levels with a more uniform and better-characterized exposure field, was not able to confirm the findings of the Repacholi study.

Some recent reviews of the field have tended to ignore the past bulk of literature on confirmed effects and instead focus on more recent controversial claims of low-level or "athermal" effects, particularly for ELF amplitude-modulated RF/microwave exposures, where it is claimed that the modulation frequency is important. These reviews include the review by the Royal Society of Canada [51], a report of a panel in the United Kingdom chaired by Sir William Stewart [52], and recent reports of the National Radiological Protection Board in the United Kingdom [53,54]. These claims of "athermal" effects in general are characterized by lack of replication and by the presence of artifacts. With regard to safety standards and guidelines, both the ICNIRP committee and the IEEE standards committee found that such effects were not useful. ICNIRP concluded "Overall, the literature on athermal effects of AM [Amplitude Modulated] electromagnetic fields is so complex, the validity of reported effects so poorly established, and the relevance of the effects to human health is so uncertain, that it is impossible to use this body of information as a basis for setting limits on human exposure to these fields" [40]. Similarly, after an extensive critical review of the literature, the committee that developed the IEEE C5.1-2005 standard concluded, "Despite more than 50 years of RF research, low-level biological effects have not been established. No theoretical mechanism has been established that supports the existence of any effect characterized by trivial heating other than microwave hearing. Moreover, the relevance of reported low-level effects to health remains speculative and such effects are not useful for standard setting" [38].

There are valid scientific considerations that make such claims implausible. The extensive paper by Valberg et al. [55] has shown that claims of low-level mechanisms are implausible at low frequencies. Furthermore, recent work by Prohofsky [56] shows that RF energy at frequencies below 300 GHz is rapidly thermalized in biological tissue (e.g., picoseconds to nanoseconds) thus making athermal effects very unlikely. It is worthwhile to recall that similar claims of "specific," rather than "thermal," effects were prevalent during the first half of the twentieth century. The challenge presented in the review by Bierman [57] and Mortimer and Osborne [58] in the 1940s remains useful today as it was then, namely, the burden of proof remains on those who claim other than heating as a mechanism for observed microwave bioeffects. One is reminded of this principle as claims of athermal effect continue to be raised [59] and rebutted [60].

21.3 Standards for the Safe Use of Microwave Energy

It is useful, at the outset, to define some terms prevalent in the standards world, especially that of the IEEE. There are three levels of authority in IEEE standards documents [61]. First, there are *standards* that imply mandatory actions and rules. The word *shall* is used often. Second, there are *recommended practices*. These describe preferred actions and rules. The word *should* is used often. Finally, there are *guides*. In these documents, alternative actions, procedures, policies, and rules are discussed. The word *may* is used and the choice of various options is at the discretion of the user.

In addition, it is useful to classify standards by their differing purpose or target for control. The most fundamental type of standard is the *exposure* standard, or safety standard. This sets limits on safe exposure of people in terms of easily measurable quantities such as field strength, power density, and induced and contact current. These limits, usually called maximum permissible exposure (MPE) levels or investigation levels, are expressed as a function of time (exposure duration). The MPEs are derived from the limiting dosimetric quantities, called basic restrictions, such as SAR, energy density (SA), and current density. At frequencies above a few GHz, the depth of penetration is only a few millimeters, or less, and the basic restrictions are expressed in terms of incident power density and are identical to the MPEs. The key feature of *exposure* standards is that they are rules that apply to *people* and it is implied to some extent that exposure is voluntary or at least the subject is aware of the exposure or acknowledges the exposure. In addition to specifying exposure limits, the IEEE C95.1 standards also include definitions and rules necessary for their implementation. Other organizations may apply different names to their documents like *guidelines* and usually do not have the strict definitions, background, and guidance for their implementation that an IEEE standard has.

A different type of standard is the *product performance* standard. This applies to a product and it is designed to ensure that potential exposures of people who use the product will be well below the basic restrictions and MPEs found in exposure standards. Examples include the laser standards and the microwave oven standard. The laser standards, for example, ANSI Z136.1 [62] and IEC 60825 [63], define accessible emission limits per stated class of laser where the classes are based on potential hazard. The microwave oven standard limits the emission of microwave energy as measured at any point 5 cm from the external surface of the oven.

Although not commonly used today, there is another type of standard that may find greater use in the future. This is the *environmental* standard in which limits are set for electromagnetic (EM) fields in the environment without a specified dependence on an exposure time, except for possible exemptions at sites where passage is transient—for example, on bridges, and so forth. The basis for an *environmental* standard would be considerations of possible interference effects, side effects like arcing at ends of long cables near a transmitting tower, and possible effects on local flora and fauna—for example, endangered bird species [64]. In addition, consideration would be made of the psychosocial factors deriving from the concerns of the people who live or work in the environment in question.

Because interference with electronic devices and equipment occurs at far lower EM field levels than biological exposure phenomena, there is also the possibility that exposures to EM fields could lead to *hazardous RFI*, for example, with medical devices like the implanted pacemaker. Although limiting exposure could control this, it is more logical to control such hazards by limiting the *susceptibility* of medical devices. In practice, often there are applied control measures on both the "radiator" and the "victim" even though the preferred (e.g., by the FCC) emphasis should be on limiting susceptibility of the "victim" device or system.

Another class of standard document may be that for prevention of accidental ignition of fuels or EEDs. This may take the form of limiting the approach or providing safe distances from various types of transmitting antennas when using electric blasting caps during explosive operations, for example, IEEE Std C95.4-2002 [65].

The history of development of microwave exposure standards has been recently presented [66,67] by the present authors, especially that of the development of the ANSI/IEEE series of C95 standards. In the early 1990s, a survey by Petersen [68] showed that exposure limits in the microwave frequency range of various organizations around the world were not greatly different (i.e., within a few dB). In the United States, besides the IEEE, safety limits for microwave exposure are also produced by the NCRP (National Council on Radiation Protection and Measurement) [69] as well as the ACGIH (American Conference of Governmental and Industrial Hygienists) [70]. In Europe, safety documents have been produced by various organizations including ICNIRP and CENELEC. In recent years European countries, besides developing national safety, documents have made increasing reference to guidelines published by ICNIRP. The latest ICNIRP guidelines [40] are now the norm for the European Union (EU) member states. The basic restrictions of the ICNIRP guidelines and the IEEE C95.1-2005 standard are identical; the MPEs of the lower tier of the ICNIRP guidelines [40] and the lower tier of the IEEE C95.1 standard [38] are the same across the frequency range of 30–100 GHz but the ICNIRP guidelines are far more conservative at low frequencies. Extensive criticism [71] has been leveled at ICNIRP for the excessive safety factor at low frequencies. The lower tier of the ICNIRP guidelines applies to the general public; the lower tier of the IEEE C95.1 standard is an action level that triggers mitigative measures, such as the implementation of an RF safety program, for example, IEEE Standard C95.7-2005 [72], to prevent exposures above the upper tier. Finally, the IEC (International Electrotechnical Commission) also develops standards relating to the safety of RF energy. For example, IEC TC-106 develops basic standards and product standards for the assessment of human exposure to electric, magnetic, and electromagnetic fields. Standards developed by TC106 describe measurement and computational techniques for assessing exposure but do not recommend safety limits. In principle, such standards can be used to assess compliance with any science-based limits.

In this review, we discuss in some detail the requirements in the latest C95.1 standard (IEEE C95.1-2005) [38] developed by the IEEE International Committee on Electromagnetic Safety (ICES) Technical Committee 95 (TC-95)—formerly IEEE Standards Coordinating Committee 28 (SCC-28). This standard was developed under the due process rules of the IEEE Standards Association with extensive documentation of all deliberations and with procedures assuring a broad consensus. At present, several hundred people of all disciplines are involved in the various subgroups of ICES with about 35% non-U.S. participation, a figure that is growing rapidly.

IEEE C95.1-2005, "IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz," represents the culmination of a 14-year effort. This standard, a major revision of IEEE Std C95.1-1991 [73], is the fourth revision of the original C95.1 standard C95.1-1966 [74]. As in the case of the earlier C95.1 standards, the 2005 revision "protects against established adverse health effects in human beings associated with exposure to electric, magnetic, and electromagnetic fields in the frequency range of 3 kHz to 300 GHz" [38]. As stated in the definitions clause of the standard "Adverse effects do not include biological effects without a harmful health effect, changes in subjective feelings of well-being that are a result of anxiety about RF effects or impacts of RF infrastructure that are not physically related to RF emissions, or indirect effects caused by electromagnetic interference with electronic devices" and "Sensations (perceptions by human sense organs) per se are not considered adverse effects." The committee concluded that speculative effects that may or may not exist (not established adverse health effects) are not useful for standard setting, which is also the conclusion of ICNIRP [40].

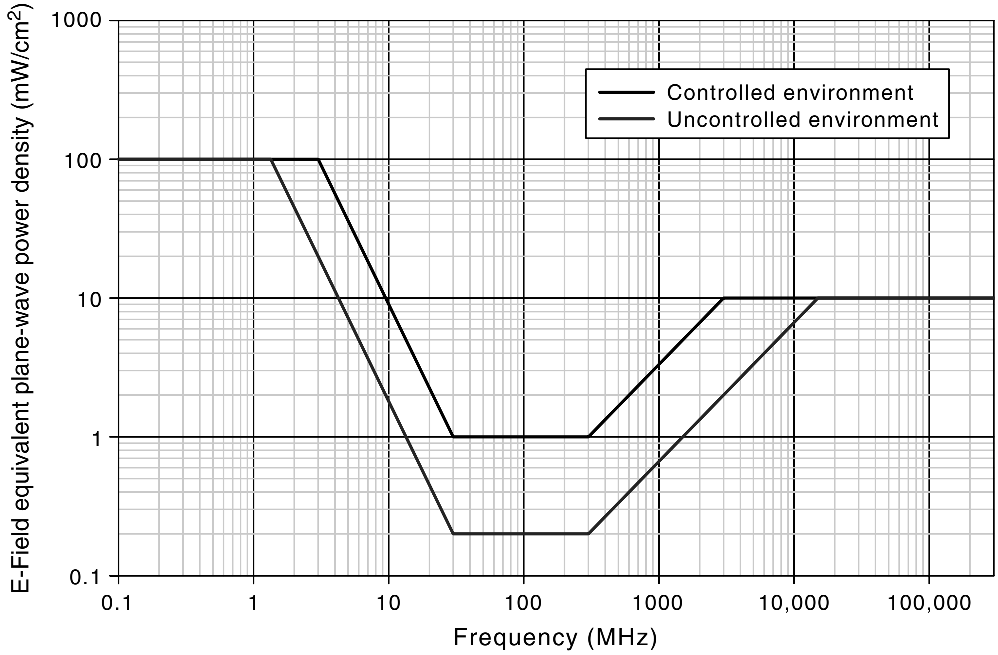
The literature database that was reviewed consisted of more than 1100 papers published through 2003, but a number of papers published after 2003 relating to mobile telephones are also included. A detailed description of the results of the review, categorized by biological endpoint, is contained in Annex B of the standard; an earlier review can be found in Supplement 6 of *Bioelectromagnetics* [39].

The C95.1-2005 standard specifies basic restrictions and MPEs over a range of frequencies extending from 3 kHz to 300 GHz. The basic restrictions for frequencies between 100 kHz and 3 GHz are expressed in terms of SAR—both whole-body-averaged and peak spatial average. From these are derived the MPEs, expressed in terms of more easily determined quantities, for example, incident field strength and power density. Above 3 GHz, both the MPEs and basic restrictions are expressed in terms of incident power density and both are equal.

As suggested by the exposure diagram of Figure 21.2, exposure limits involve an *averaging time.* Thus when comparing exposure to the MPEs or basic restrictions, the measured or calculated exposure values are time- averaged over any contiguous period of time equal to the averaging time. This means that higher exposures are permitted for short periods of time as long as the time average is below the corresponding MPE.

The basic rules of C95.1 are presented in Figure 21.3a and b and Figure 21.4. In Figure 21.3a, the C95.1-1991 MPEs for the lower tier (exposures in uncontrolled environments) and for the upper tier (exposures in controlled environments) are shown in terms of the E-field plain-wave equivalent power density as a function of frequency. In Figure 21.3b, the C95.1-2005 MPEs for the lower tier (action level) and for the upper tier (for exposures in controlled environments) are shown; the lower tier of the ICNIRP MPEs are also shown. It is important to note that the MPEs of the lower tier of the 1991 C95.1 standard (Figure 21.3a) were changed in 2005 (Figure 21.3b) for purposes of global harmonization and social

(a)



(b)

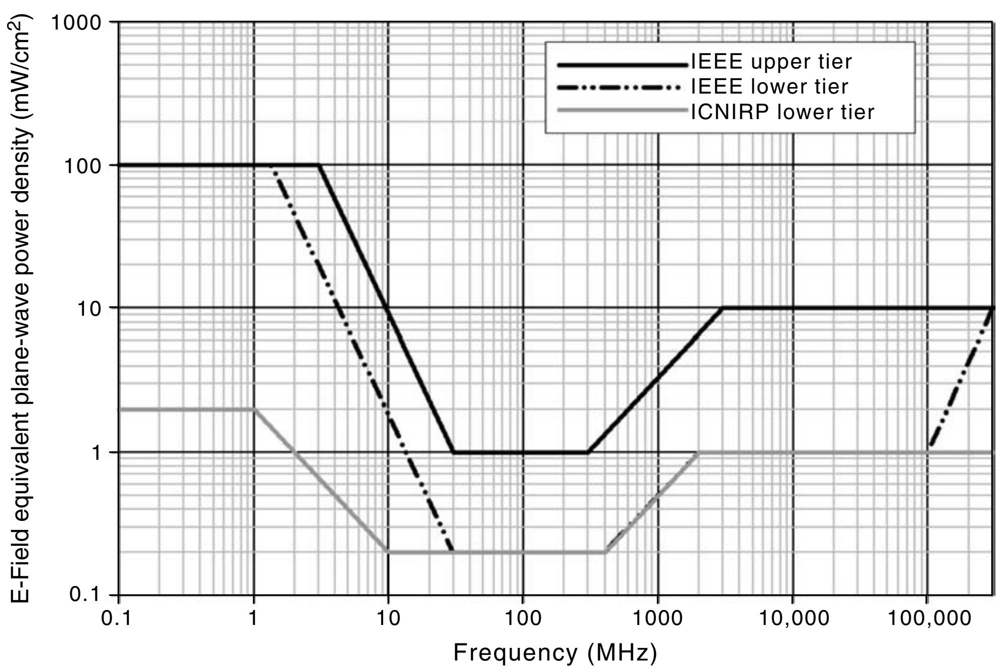


FIGURE 21.3 (a) The IEEE C95.1-1991 MPEs (in terms of the E-field plane-wave equivalent power density). (b) The IEEE C95.1-2005 MPEs (in terms of the E-field plane-wave equivalent power density). The corresponding ICNIRP MPEs [40] for the lower tier, which are the same as the corresponding IEEE MPEs between 30 MHz and 100 GHz, are shown for purposes of comparison.

policy—the upper tier is still considered safe for all. In Figure 21.4, a "capsule" presentation of most of the features of the C95.1 standard is shown. Some key observations are as follows:

1. The use of two "tiers." The lower tier, or action level," serves as a trigger at which point, mitigative measures should be applied to preclude exposures that exceed the upper tier. This generally applies when the exposed population is unaware of their exposure and is not competent to address safety measures. The necessity for a lower tier is debatable as it is clearly stated in IEEE C95.1-2005 that "The upper tier, which is protective for all with an acceptable margin of safety, applies to exposure of persons in controlled environments. While the weight of scientific evidence supports the conclusion that there is no measurable risk associated with RF exposures below the upper tier of this standard, it is scientifically impossible to prove absolute safety (the null hypothesis) of any physical agent. Thus the lower tier, with an additional safety factor, recognizes public concerns and also supports the process of harmonization with other standards . . ." [38]. The use of two tiers originated [75] in an attempt to apply an exposure standard conservatively to the environment. In the future, the lower tier may disappear from standards like C95.1.
2. The use of the SAR and SA concepts as basic restrictions applies only between 100 kHz and 3 GHz (6 GHz in C95.1-1991 [73] and 10 GHz in the ICNIRP guidelines [40]), in which range there are significant internal fields from RF exposure. Below 100 kHz, the limits derive from electrostimulation phenomena. Above 3 GHz, there is little penetration (e.g., less than a few millimeters) into the body and surface absorption determines bioeffects and hazards. Because of these facts, the MPEs in the 1991 C95.1 standard [73] were raised to 10 mW/cm^2 above 3 and 15 GHz, respectively, for the controlled and uncontrolled environments (see Figure 21.3a). The corresponding MPEs for the lower tier were lowered in the 2005 revision for "social policy" reasons and to more easily move toward international harmonization [38] (see Figure 21.3b).
3. Averaging times are ramped with frequency in accordance with the thermal time constants of the human body. In the 1990s, two intensive conferences [76] between the laser and microwave communities were held with the goal of improving the agreement between laser and microwave standards in the transition range around 300 GHz. From these conferences and further study, a more representative ramp in averaging time to better fit the dependence of thermal time constants versus frequency was developed, which is included in the 2005 revision of the C95.1 standard.
4. In response to a request from the IEEE MTT Society, a limit on peak exposure levels for very short pulses was first introduced in the 1991 C95.1 standard. It consists of a value of 100 kV/m maximum as well as an energy limit for short pulses. These are based on studies by the U.S. Air Force that showed that the potential hazards of short pulses were of the "stun" variety and associated with thermal phenomena. The peak limits are well above theoretical thresholds for the auditory effect.
5. Limits on induced or contact currents are included but these are continually being expanded and refined to include spark-discharge phenomena as well as point contact currents. Also included are values of the electric field strength below which induced current does not have to be measured.
6. Relaxation of limits for localized (partial-body) exposure was included in an incomplete and tentative basis in the 1991 standard, which included a caveat on exposure of the testes and eyes. The caveat has been eliminated in the 2005 revision because of the improved ramps for averaging time at high frequencies that eliminate undesirable high-energy content in short duration exposures.
7. The basic restrictions for the upper tier in terms of SAR are 0.4 W/kg whole-body-average and 8 W/kg peak spatial average. They first appeared in the 1982 ANSI C95.1 standard [77]. These were derived from far-field animal exposure studies associated with behavioral disruption. The 8 W/kg (in any 1 g of tissue in the shape of a cube) peak spatial average derives from the 20:1 peak to average ratio of the SAR distributions observed in experimental animals under far-field exposure conditions; a tissue mass of 1 g over which the spatial peak SAR value is averaged was considered the approximate resolution of contemporary dosimetry techniques at the time. In C95.1-1991 and C95.1-2005, these have been extended in a doubly conservative manner to the control of localized exposures from sources like mobile phones, first by invoking the lower tier reduction

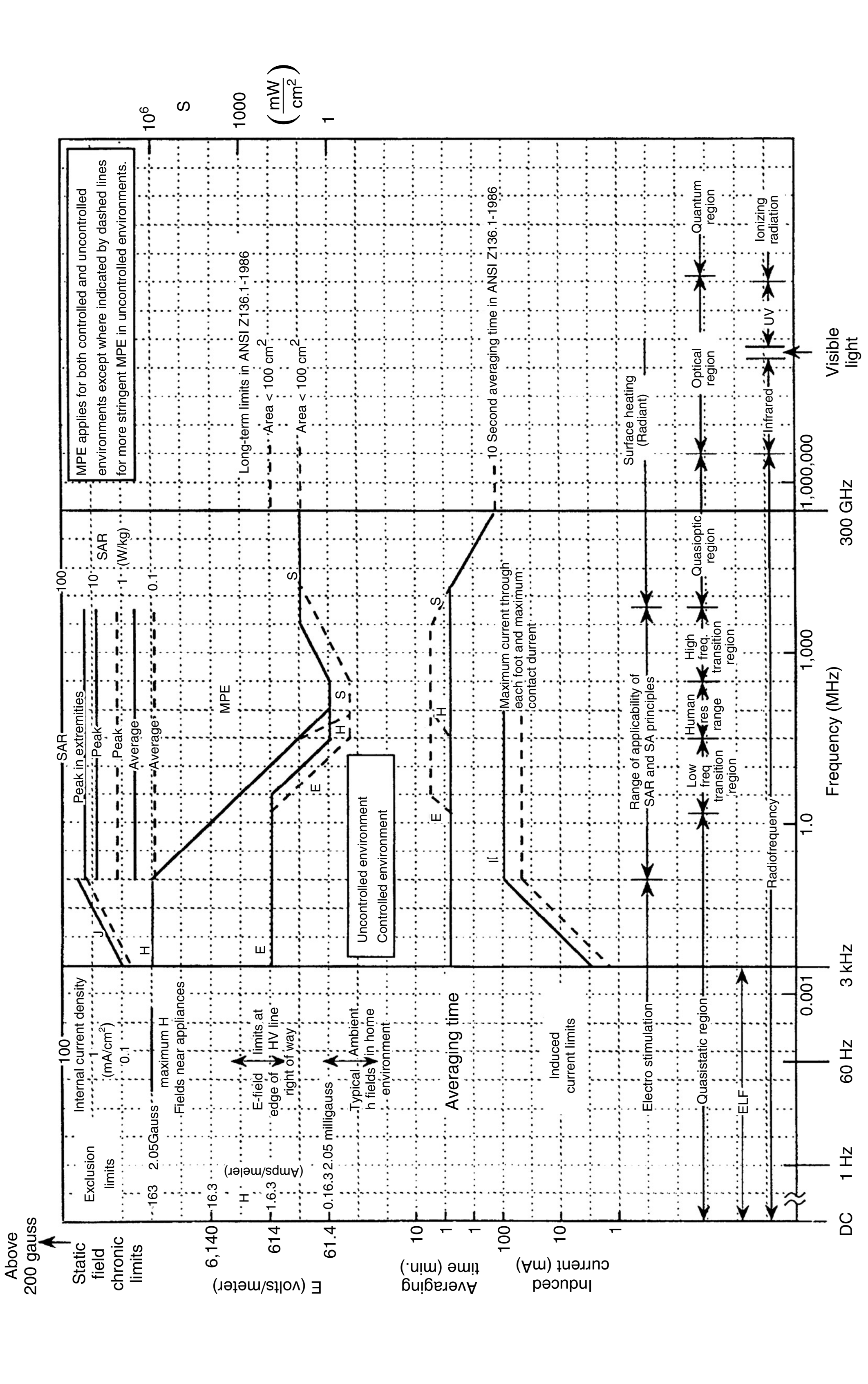


FIGURE 21.4 A capsule guide to the IEEE C95.1-1991 standard showing all of the relevant features. Although the lower tier has been modified slightly above 300 MHz, the guide brings together the basic concepts.

> of 5 and then by assuming that the spatial peak from whole-body experiments should apply to partial-body exposure. For purposes of harmonization with other international guidelines, the peak spatial average SAR has been relaxed in the C95.1-2005 standard from 8 and 1.6 W/kg for the upper and lower tiers found in the 1991 standard (both averaged over any 1 g of tissue) to 10 and 2 W/kg averaged over any 10 g of tissue. These relaxations are supported by calculations and measurements that have been made using tissue models for the case of exposures of the head near small antennas operating in the 900 MHz or 1.8 GHz bands (representative of exposure from handheld mobile telephones), for example, Wang and Fujiwara [78], Riu and Foster [79], and Van Leeuwen et al. [80].

Besides C95.1, important C95 standards and recommended practices include a recently issued reaffirmation IEEE C95.2 (RF warning symbols) [81], the revision of the C95.3 (RF measurements and computations) [82], a revision of IEEE C95.4-2002 (safe distances from antennas during blasting operations) [65], and a new standard C95.7-2005 (RF safety programs) [72]. In addition, a new standard, IEEE C95.6-2002, which provides basic restrictions and MPEs for exposures to electric and magnetic fields at frequencies between 0 and 3 kHz was approved in 2002 [83].

Product performance standards for the microwave technologies arose after the passage of the Radiation Control for Health and Safety Act of 1968 (P.L. 90-602) [84]. This led to the emission standard [85] for microwave ovens and to some control of industrial and other microwave sources. All the currently applicable regulations are accessible at Internet site http://www.fda.gov/cdrh/radhlth/index.html [86]. It is unlikely that oven limits will be changed for health reasons. Because of an impending RFI scenario [87] involving microwave ovens and wireless applications operating in the oven (ISM) band, there are some pressures to voluntarily limit emissions from ISM equipment and even encourage the FCC to narrow that ISM band.

In the United States, most civilian applications of microwave/RF energy are now regulated by FCC Rules [88], which are based in large part on the C95.1 standards and also on the NCRP guidelines at higher microwave frequencies. The FCC is concerned not only with regulation of the environment around transmitting towers and antennas but also with the exposure to the user of wireless handsets. In this sense, the FCC has taken on the responsibility that normally under P.L. 90-602 would fall to the Food and Drug Administration (FDA). In any case, the FCC as well as the FDA supports the derivation of product standards through an open consensus process such as that of the IEEE. Although FDA has the authority to develop performance standards for all microwave equipment including wireless phones, it has instead supported the creation of a committee to develop such standards, namely, IEEE Standards Coordinating Committee 34.

IEEE SCC-34 (now ICES Technical Committee 34) is a relatively new committee having been established in 1995 (compared with ICES TC-95, which was first established as an American Standards Association Committee in 1960) for the purpose of developing product performance standards relative to the safe use of electromagnetic energy. The committee uses the exposure criteria and basic restrictions developed by TC-95, and in some cases, by other committees, to develop standardized assessment procedures, emission limits, and so forth, to allow manufacturers to readily ensure that their products comply with these criteria. The goal is to develop unambiguous procedures that yield repeatable results, for example, similar to the procedure for certifying compliance of microwave ovens. So far, three subcommittees have been established. The subjects being addressed are pleasure boat radar, protocols for assessing the efficacy of RF protective garments, and protocols for certifying compliance of wireless handsets, vehicle-mounted antennas, and other similar devices using both measurement and numerical techniques.

Protocols for assessing the peak spatial average SAR associated with the use of handheld radio transceivers used for personal communications, for example, cell phones, are especially important as there remains a lingering, albeit unsupportable, concern by some members of the public. This concern first arose in 1993 when a guest on a TV talk show alleged that his wife's brain tumor was exacerbated by the use of a cell phone. Although the health issue continues to dwindle, the media continues to raise concern by bringing inordinate attention to preliminary results of every study reported that even suggests an association between untoward medical effects and the use of cell phones—even when the exposure paradigm is completely different. During the past 5 or 6 years, attention has focused on differences between SAR measurement results reported by different laboratories for the same phone. By a major leap of logic, small differences, which are not unexpected in light of the different protocols being used to test cell phones, are translated to a theme of uncertainty about cell phone safety. This seems to occur more in this field than in many others, that is, a focus on uncertainty related to small differences in analytical or measurement results while completely ignoring the issue of how far below established safety (exposure) criteria the results may be.

Following the 1993 program that first raised the safety issue, a 25–27 million US dollar research program was established. This program, funded by industry through a blind trust, was to review the literature, develop a research agenda, and fund the studies necessary to address criticisms levied against judgments based on interpretations of the extant literature. One such criticism, unwarranted as it may be, was that there was a lack of studies at exactly the same frequencies, modulations, and exposures as those associated with cell phones. Part of the research program was the establishment of a dosimetry working group to develop uniform protocols for assessing exposure from wireless handsets. When funding was withdrawn for this particular project, the working group, which by then included representatives from most handset manufacturers, a number of test houses and academia, evolved into Subcommittee 2 of ICES TC-34.

Two separate working groups were established to develop recommended practices—one addresses experimental techniques (measurements) and the other addresses numerical techniques. The experimental technique utilizes robot-controlled miniature electric field probes to scan and measure the SAR in a homogeneous tissue-simulating liquid-filled anthropomorphic model of the human head. Although there are a number of head models available for purposes of certifying handheld mobile phones using measurement techniques, the Specific Anthropomorphic Model (SAM) has become the industry standard and is defined in standards such as IEEE 1528-2003 [89], 1528a-2005 [90], and IEC 62209-2004 [91]. The numerical technique applies the Finite Difference Time Domain (FDTD) method to solving Maxwell's equations in a heterogeneous representation of the human head developed from CT and MRI scans of humans. Models with resolutions of $2 \times 2 \times 2$ mm [92], $1.1 \times 1.1 \times 1.4$ mm using subgridding in some regions [93], and $0.9 \times 0.9 \times 1.5$ mm [94] have been reported.

Both techniques have advantages and disadvantages. For example, the experimental technique uses the actual cell phone as the source but the homogeneous head model is less than an ideal representation. Head size, the dielectric properties of the tissue simulant, and the thickness of the spacer representing the pinna are standardized to represent a worst case situation, which overestimates the SAR induced in the brain. With the numerical technique, the head model is an accurate representation of a human head but the handset has to be modeled, usually as a simple metal box with an appropriate antenna. CAD files of actual phones complete with some internal structures have been used and differences between these and the results from the simple model continue to be investigated. An advantage of the computational technique is that it can be applied at the design stage to optimize antenna performance while ensuring that the peak SAR is below the specified limit.

Because most manufacturers use the experimental technique to certify their products, mainly because measurement systems are available commercially, the initial effort was directed to completing this standard. At the time the standard was developed, much of the information needed to complete the document was not available in the literature. For example, assessment of the uncertainty associated with each component of the system and the overall assessment uncertainty was developed in the laboratories of the committee members as the standard developed. This included a series of interlaboratory comparisons of canonical models such as standard half-wave dipoles above a flat phantom or sphere and the development of a generic phone by three of the manufacturers for further interlaboratory comparisons. The committees developing these particular standards, for example, CENELEC, IEC, IEEE, have many members in common to ensure international harmonization. Both the FDA and the FCC play an active role on IEEE TC-34.

The results of these efforts led to IEEE Standard 1528-2003 [89], which predated IEC Standard IEC62209 [91]—but the two are in complete harmony with each other and with similar standards such as the European Standard (CENELEC) EN50361 [95]. Future projects will be the extension of the handset measurement protocols to other wireless devices, for example, wireless modems, body-mounted radio transceivers, and the use of numerical techniques to determine the SAR associated with exposure to these and other transmitting devices, for example vehicle-mounted antennas. Having standard protocols in place should mitigate some of the media-driven concern about wireless devices exemplified by TV "specials" that, for example, call attention to the uncertainty of cell phone safety by focusing on the uncertainty of the peak SAR, no matter how far below the basic restrictions.

In recent years, there is a trend toward international harmonization of standards in the microwave/RF area. ICES TC-34 and TC-95 are very involved in this activity both by expansion of non-U.S. membership but also by closer liaison with other international groups such as the IEC and ICNIRP. In this global activity, it is the intent of the ICES to ensure that the benefits of the IEEE standards process with due process and broad consensus of all stakeholders are recognized. The importance of rational science-based standards cannot be overstated. Excessive safety factors, as in the ICNIRP'98 guidelines, can unduly suppress technology, for example, antitheft and weapons detector systems. In addition, uneven safety factors across the spectrum can give unfair advantage to one technology versus another that uses a different part of the spectrum and, almost as important, convey a sense of uncertainty about the science. And, too conservative safety limits can lead to inferior performance and loss of benefit and reliability. This could be important if it involves emergency communications, for example.

21.4 Risk Assessment and Public Education

The fears of microwaves along with a generalized electrophobia were promoted in large measure by the writings of Paul Brodeur [96]. In succession, there have been waves of fear targeted at microwave ovens, radars, microwave relay towers, visual-display terminals, power lines, electric blankets, police radar, cellular phones, and wireless base stations. More recently, there arose a claim that the PAVE PAWS radar was uniquely dangerous because of peculiar features ("steep wavefronts") in the radiated RF energy. An extensive study by the National Research Council, however, found that the theoretical speculations had no scientific basis and that there was no evidence of hazard from that radar in the environment [97].

A historian, Nicholas Steneck [98], in the 1980s attacked the C95 standards as being too science-based. He felt exposure limits should be as low as possible and not based on actual hazard thresholds. When the power line (50 Hz) controversy erupted in the 1990s, Morgan [99] proposed the practice of "prudent avoidance," just in case the weak allegations of a cancer link were true. As documented by Park [100], however, the power line scare was baseless and derived from "voodoo science."

Now, in the context of a mushrooming global wireless technology, a plethora of proposals to apply the same caution with a new name has arisen, namely, the Precautionary Principle [101]. It has been endorsed by the Stewart panel [52] in the United Kingdom for use in processing applications for wireless base stations. It was also endorsed by the European Commission [102]. Despite the new name, this proposal is simply another tool for attacking reliance on science-based standards. In response to a request from members of a federal Interagency Committee charged with writing a report at the conclusion of the RAPID program on EMF research, IEEE ICES (SCC-28) submitted advice [103] that discussions of cautionary polices and procedures should be done at the lowest level of authority—that is, in preparing *guides* that present optional alternatives that individuals or organizations can adopt on an ad hoc discretionary basis. In this way, there is no acknowledgment of such a cautionary principle as avoiding or superseding reliance on standards. In the United States, the precautionary principle has not been endorsed by the federal government [104].

This new principle simply adds to the forces of electrophobia. It raises the level of fear of microwaves and stimulates the sale of a wide variety of allegedly protective devices and services. This amounts to an exploitation of electrophobia and in general is to be discouraged. These devices include measurement devices, shields, books, newsletters, and so forth. Besides selling oven leakage detectors, discouraged by both the FDA and leading consumer organizations, the exploiting parties give advice on how to minimize exposure—for example, by "running' out of the kitchen after turning on the microwave oven. The IEEE

COMAR committee [105] continues to fight this phobia and allied misinformation through a series of Technical Information Statements. Other organizations also provide useful information.

A partial list of useful Web sites includes

CTIA:	http://www.WOW-COM.com
FCC:	http://www.fcc.gov/oet.rfsafety/cellpcs.html
FDA:	http://www.fda.gov/cdrh/radhlth/index.html
ICNIRP:	http://www.icnirp.de
IEEE/COMAR:	http://ewh.ieee.org/soc/embs/comar
IMPI:	http://www.impi.org
IEE ICES TC95 IEEE ICES:	http://grouper.IEEE.org/groups/scc28
IEEE ICES TC-34	http://grouper.IEEE.org/groups/scc34
Dr. Moulder/FAQ:	http://www.mcw.edu/grcr/cop.html
WHO:	http://www.who.ch/
Health Council of the Netherlands:	http://www.gr.nl/index.php?phpLang=en
Health Canada:	http://www.hc-sc.gc.ca/iyh-vsv/prod/index_e.html
National Radiological Protection Board:	http://www.nrpb.org/publications/documents_of_nrpb/abstracts/absd15-3.htm
Nordic Authorities:	http://www.ssi.se/english/english_news.html
ARPANSA:	http://www.arpansa.gov.au/pubs/eme_comitee/fact1.pdf

These are some of the Web sites where information useful to microwave engineers can be obtained. The FDA Web site also links to an educational site maintained by a professor at the University of Virginia. It contains much useful information but errs in recommending the use of inexpensive oven leakage detectors and presenting a less than complete explanation of superheating of liquids in microwave ovens. On the Web site for the FCC, much detailed information is available including a series of reports, fact sheets, and an interactive database containing the dielectric properties for a number of biological tissues.

The site for IMPI (International Microwave Power Institute) can be useful in exploring power (non-communications) applications of microwave energy. These include medical applications, food heating and processing, microwave discharge lamps for ultraviolet (UV) curing and lighting, and microwave power transmission, some of it for solar power satellites. Other future applications include the comfort heating of animals and human through microwaves that can contribute toward solving energy crises. The other sites are self-explanatory in their functions.

In the future, there will be emerging interest in the "terahertz" spectrum [106] and the deployment of microwave weapons [107,108]. These subjects will pose new challenges for standards setting as well as public education with rational public policy as the goal.

21.5 Conclusions

The modern microwave engineer needs to be knowledgeable about safety standards in general, and in particular, the IEEE C95 series of standards. Microwave engineers through their professional societies as well as employers should support and participate in the increasingly global activities of the IEEE in developing standards for the safe use of electromagnetic energy. The financial support of all IEEE activities in global standards and public education will be critical in the future development of microwave technology. With rational standards and an educated public, all microwave technologies, both for communications and power applications, can be realized for optimum benefit to mankind.

References

1. Reilly, J. P., *Applied Bioelectricity: From Electrical Stimulation to Electropathology*, Springer Verlag, New York, 1998.

2. Reilly, J. P., "Mechanisms of electrostimulation: application to electromagnetic field exposure standards at frequencies below 100 kHz." In: P. Chadwick and C. Gabriel (eds.), *The International EMF Dosimetry Handbook*, Published as an Internet book available at Web site http://www.emfdosimetry.org
3. Durney, C. H., Massoudi, H., and Iskander, M. F., *Radiofrequency Radiation Dosimetry Handbook*, Fourth Edition, Report USAFSAM-TR-85-73, USAF School of Aerospace Medicine, Brooks AFB, TX, 1986.
4. Gabriel, S., Lau, R. W., and Gabriel, C., "The dielectric properties of biological tissues: III Parametric models for the dielectric spectrum of tissues," *Phys. Med. Biol.*, vol. 41, pp. 2271–2293, 1996. These properties can also be downloaded from Web site: http://www.brooks.af.mil/AFRL/HED/hedr/reports/dielectric/home.html
5. Schwan, H. P., "History of the genesis and development of the study of low energy electromagnetic fields," In M. Grandolfo, S. Michaelson, and A. Rindi, (eds.), *Biological Effects and Dosimetry of Non-ionizing Radiation*, pp. 1–17. Plenum Publishing Company, New York, 1981.
6. Osepchuk, J. M., "Basic characteristics of microwaves," *Bull. N. Y. Acad. Med.*, vol. 55, pp. 976–998, 1979.
7. Chou, C. K., Bassen, H., Osepchuk, J., Balzano, Q., Petersen R., Meltz, M., Cleveland, R., Lin, J. C., and Heynick, L., "Radio frequency electromagnetic exposure: tutorial review on experimental dosimetry," *Bioelectromagnetics*, vol. 17, pp. 195–206, 1996.
8. Kritikos, H. N. and Schwan, H. P., "The distribution of heating potential inside lossy spheres," *IEEE Trans. Bio. Med. Eng.*, vol. BME-22, pp. 457–463, 1975.
9. Taflove, A., *Advances in Computational Electrodynamics*, Artech House, Norwood, MA, 1998.
10. USAF, "2002 EMF Dosimetry Models," United States Air Force. Available at Web site: http://www.starview.brooks.af.mil/EMF/dosimetry_models
11. Mason, P. A., Hurt, W. D., Walters, T. J., D'Andrea, J. A., Gajsek, P., Ryan, K. L., Nelson, D. A., Smith, K. I., and Ziriax, J. M., "Effects of frequency, permittivity, and voxel size on predicted specific absorption rate values in biological tissue during electromagnetic-field exposure," *IEEE Trans. Microwave Theory Techniques*, vol. 48, pp. 2050–2058, 2000.
12. Dimbylow, P. J., "Fine resolution calculations of SAR in the human body for frequencies up to 3 GHz," *Phys. Med. Biol.*, vol. 47, pp. 2835–2846, 2002.
13. Nikoloski, N., Frohlich, J., Samaras, T., Schuderer, J., and Kuster, N., "Reevaluation and improved design of the TEM cell in vitro exposure unit for replication studies," *Bioelectromagnetics*, vol. 26, pp. 215–224, 2005.
14. Schonborn, F., Pokovic, K., Burkhardt, M., and Kuster, N., "Basis for optimization of in vitro exposure apparatus for health hazard evaluations of mobile communications," *Bioelectromagnetics*, vol. 22, pp. 547–559, 2001.
15. Guy, A. W., Lehmann, J. F., and Stonebridge, J. B., "Therapeutic applications of microwave power," *Proc. IEEE*, vol. 62, pp. 55–75, 1975.
16. Durney, C. H. and Christensen, D. A. "Hyperthermia for cancer therapy, "in Gandhi, O. P. (ed.), *Biological Effects and Medical Applications of Electromagnetic Energy*, Prentice-Hall, Englewood Cliffs, NJ, pp. 436–477, 1990.
17. Moritz, A. R. and Henriques, F. C., Jr., "Studies of thermal injury II: The relative importance of time and surface temperature in the causation of cutaneous burns," *Am. J. Pathol.*, vol. 23, pp. 695–720, 1947.
18. Michaelson, S. M., Thomson, R. A. E., and Howland, J. W., "Physiologic aspects of microwave irradiation of mammals," *Amer. J. Physiol.*, vol. 201, pp. 351, 1961.
19. Addington, C. H., Osborn, C., Swartz, G., Fischer, F., Neubauer, R. A., and Sarkees, Y. T., "Biological effects of microwave energy at 200 megacycles," In M. F. Peyton, (ed.), *Proc. 4th Annual Tri-Service Conf. Biol. Effects of Microwave Radiation*, Plenum Press, NY, p. 177, 1961.
20. Minin, B. A., *Microwaves and Human Safety*, I. Translation from U. S. Joint Publications Research Service, JPRS 65506-1, U. S. Dept. of Commerce, 1965.
21. Pay, T. L., "Microwave effects on reproductive capacity and genetic transmission in *Drosophila melanogaster*," *J. Microwave Power*, vol. 7, pp. 175, 1972.
22. Samaras, G. M., Muroff, L. R., and Anderson, G. E., "Prolongation of life during high-intensity microwave exposures," *IEEE Trans. Microwave Theory Techniques*, vol. MTT-19(2), pp. 245–247, 1971.
23. Tell, R. A. and Harlen, F., "A review of selected biological effects and dosimetric data useful for development of radiofrequency safety standards for human exposure," *J. Microwave Power*, vol. 14(4), pp. 405–424, 1979.

24. Adair, E. R., Mylacraine, K. S., and Kelleher, S. A., "Human exposure at two radiofrequencies (450 and 2450 MHz): similarities and differences in physiological response," *Bioelectromagnetics*, vol. 20, pp. 12–20, 1999.
25. Adair E. R., Mylacraine K. S., and Allen S. J., "Thermophysiological consequences of whole body resonant RF exposure (100 MHz) in human volunteers," *Bioelectromagnetics*, vol. 24, pp. 489–501, 2003.
26. Adair E. R., Blick D. W., Allen S. J., Mylacraine K. S., Ziriax J. M., and Scholl, D. M., "Thermophysiological responses of human volunteers to whole body RF exposure at 220 MHZ," *Bioelectromagnetics*, vol. 26 (Suppl. 7), pp. 448–461, 2005.
27. Carpenter, R. L., "Ocular effects of microwave radiation," *Bull. N.Y. Acad. Med.*, vol. 55, pp. 1048–1057, 1979.
28. Guy, A. W., Kramar, P. O., Harris, C. A., and Chou, C. K., "Long-term 2450 MHz CW microwave irradiation of rabbits: methodology and evaluation of ocular and physiologic effects," *J. Microwave Power*, vol. 15, pp. 37–44, 1980.
29. Kues, H. A., Hirst, L. W., Lutty, G. A., D'Anna, S. A., and Dunkelberger, G. R., "Effects of 2.34 GHz microwaves on primate corneal endothelium," *Bioelectromagnetics*, vol. 6, pp. 177–188, 1985.
30. Kamimura, Y., Saito, K.-I., Saiga, T., and Amenmiya, Y., "Effect of 2.45 GHz microwave irradiation on monkey eyes," *IEICE Trans. Commun.*, vol. E77-B, pp. 762–764, 1994.
31. Osepchuk, J. M. (ed.), *Biological Effects of Electromagnetic Radiation*, John Wiley & Sons, NY, 1983.
32. Gandhi, O. P. (ed.), "Special Issue on Biological Effects and Medical Applications of Electromagnetic Energy", *Proc. IEEE*, vol. 68, 1980.
33. Elder, J. A. and Cahill, D. F., *Biological Effects of Radiofrequency Radiation*, Report EPA-600/8-83-026F, 1984.
34. Polson, P. and Heynick, L. N., "Overview of the radiofrequency radiation (RFR) bioeffects database," In B. J. Klauenberg, M. Grandolfo, and D. N. Erwin (eds.), *Radiofrequency Radiation Standards*, Plenum Press, NY, pp. 337–390, 1995.
35. Michaelson, S. M. and Lin, J. C., *Biological Effects and Health Implications of Radiofrequency Radiation*, Plenum Press, NY, 1987.
36. Gandhi, O. P. (ed.), *Biological Effects and Medical Applications of Electromagnetic Energy*, Prentice-Hall, Englewood Cliffs, NJ, 1990.
37. Polk, C. and Postow, E., (eds.), *CRC Handbook of Biological Effects of Electromagnetic Fields*, CRC Press, Inc., Boca Raton, FL, 1995.
38. IEEE C95.1-2005, *IEEE Standard for Safety Levels with respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz*, Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2005.
39. Greenebaum, B. (ed.), "Special issue: Reviews of the effects of RF fields on various aspects of human health," *Bioelectromagnetics*, Suppl. 6, pp. S1–S213, 2003.
40. ICNIRP (International Commission on Non-Ionizing Radiation Protection), "Guidelines for limiting exposure to time-varying electric, magnetic, and electromagnetic fields (up to 300 GHz)," *Health Phys.*, vol. 74, pp. 494–522, 1998.
41. Lin, J. C., *Microwave Auditory Effects and Applications*, Charles C. Thomas, Springfield, IL, 1978.
42. de Lorge, J. O. and Ezell, C. S., "Observing responses of rats exposed to 1.28 GHz and 5.4 GHz microwaves," *Bioelectromagnetics*, vol. 1, pp. 183–198, 1980.
43. Petersen, R. C., "Radiofrequency/microwave protection guides," *Health Phys.*, vol. 61, pp. 59–67, 1991.
44. Stevenson, Sen. A., U.S. Senate Hearing record on Oversight of Radiation health and Safety, June 16–29, 1977.
45. Guy, A. W. and Foster, K. R., "The microwave problem," *Scientific American*, vol. 255(3), pp. 32–39, 1986.
46. Foster, K. R. and Pickard, W. F., "The risks of risk research," *Nature*, vol. 330, pp. 531–532, 1987.
47. Belyaev, I. Y., Sheheglov, V. S., Alipov, Y. D., and Polunin, V. A., "Resonance effect of millimeter waves in the power range of 10^{-19} to 3×10^{-3} W/cm^2 on *Escherichia coli* cells at different concentrations," *Bioelectromagnetics*, vol. 17(4), pp. 312–321, 1996.
48. Osepchuk, J. M. and Petersen, R. C., "Comments on 'Resonance effect of millimeter waves in the power range from 10^{-19} to 3×10^{-3} W/cm^2 on *Escherichia coli* cells at different concentrations,' Belyaev et al., Bioelectromagnetics, vol. 17, pp. 312–321 (1996)," *Bioelectromagnetics*, vol. 18(7), 527–528, 1997.

49. Repacholi, M. H., Basten, A., Gebski, V., Noonan, D., Finnie, J., and Harris, A. W., "Lymphomas in E Warning: UNREGISTERED Word2TeX – Figure Ignored! -Pim 1 transgenic mice exposed to pulsed 900 MHz electromagnetic fields," *Radiat. Res.*, vol. 147, pp. 631–647, 1997.
50. Utteridge T. D., Gebski V., Finnie J. W., Vernon-Roberts B., and Kuchel T. R., "Long-term exposure of Eμ-Pim1 transgenic mice to 898.4 MHz microwaves does not increase lymphoma incidence," *Radiat. Res.*, vol. 158, pp. 357–364, 2002.
51. Royal Society of Canada Expert Panel Report, *A Review of the Potential Health Risks of Radiofrequency Fields from Wireless Telecommunication Devices*, Royal Society of Canada, RSC.EPR 99-1, 1999.
52. Independent Expert Group on Mobile Phones, Stewart, Sir William, Chairman, *Mobile Phones and Health*, NRPB, Chilton, Didcot, Oxon OX11 0RQ, UK, 2000.
53. NRPB, "Consultation Document—Proposals for limiting exposure to electromagnetic fields (0–300 GHz)," National Radiological Protection Board, Chilton, Didcot, OX11 0RQ, 185, UK, May 2003.
54. NRPB, "Advice on limiting exposure to electromagnetic fields (0–300 GHz)," *Documents of the NRPB*, vol. 15, no. 2, National Radiological Protection Board, Chilton, Didcot, Oxfordshire, UK, 2004b.
55. Valberg, P., Kavet, R., and Rafferty, C. N., "Can low-level 50/60 Hz electric and magnetic fields cause biological effects?", *Radiat. Res.*, vol. 148, pp. 2–21, 1997.
56. Prohofsky, E. W., "RF Absorption involving biological macromolecules," *Bioelectromagnetics*, vol. 25(5), pp. 441–451, 2004.
57. Bierman, W., *The Medical Applications of the Short-Wave Current*, Williams and Wilkins Co., Baltimore, 1942.
58. Mortimer, B. and Osborne, S. L., *J. Am. Med. Assoc.*, 104, 1413, 1935.
59. Blank, M and Goodman, R., "Comment: A biological guide for electromagnetic safety: The stress response," *Bioelectromagnetics*, vol. 25(6), pp. 642–646, 2004.
60. Petersen, R.C., "Reply to "A biological guide to electromagnetic safety: The stress response,' by Blank, M., Goodman, R.," *Bioelectromagnetics*, vol. 25(6), pp. 647–648, 2004.
61. IEEE, *IEEE Standards Style Manual*, Institute of Electrical and Electronics Engineers Standards Association, Piscataway, NJ, 2005.
62. ANSI Z136.1-2000, *American National Standard for the Safe Use of Lasers*, American National Standards Institute, New York, NY.
63. IEC 60825-1, *IEC Standard Safety of Laser Products—Part 1: Equipment Classification, Requirements and User's Guide*, 2003.
64. Osepchuk, J. M. "Environmental standards: the new concept and key to international harmonization of safety standards for the safe use of electromagnetic energy," *Proc. IEEE Int. Symposium on Technology and Society*, Worcester Polytechnic Institute, 2004.
65. IEEE C95.4-2002, IEEE Recommended Practice for Determining Safe Distances From Radio Frequency Transmitting Antennas When Using Electric Blasting Caps During Explosive Operations.
66. Petersen, R. C., "Radiofrequency safety standards-setting in the United States," Bersani, F. (ed.)., *Electricity and Magnetism in Biology and Medicine*, Plenum Press, NY, pp. 761–764, 1999.
67. Osepchuk, J. M. and Petersen, R. C., "Historical review of RF exposure standards and the International Committee on Electromagnetic Safety (ICES)," *Bioelectromagnetics*, vol. 24(S6), pp. S7–S16, 2003.
68. Petersen, R. C., "Radiofrequency/microwave protection guides," *Health Phys.*, vol. 61(1), July 1991.
69. NCRP, Biological Effects and Exposure Criteria for Radiofrequency Electromagnetic Fields. Bethesda: NCRP Report no. 86, National Council on Radiation Protection and Measurements, Bethesda, MD, 1986.
70. ACGIH, "2005 TLVs and BEIs Based on the Documentation of the Threshold Limit Values for Chemical Substances and Physical Agents & Biological Exposure Indices," American Conference of Governmental Industrial Hygienists, Cincinnati, OH, 2005.
71. Various authors, Series of letters commenting on ICNIRP'98 guidelines, *Health Phys.*, vol. 76–77, 1999.
72. IEEE C95.7-2005, "IEEE Recommended Practice for Radio Frequency Safety Programs," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2005.
73. IEEE C95.1-1991 "IEEE Standard for Safety Levels with Respect to Human Exposure to Radio Frequency Electromagnetic Fields, 3 kHz to 300 GHz," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 1999.
74. USAS C95.1-1966 "USA Standard Safety Level with Respect to Personnel," American Standards Association (now ANSI), New York, NY, 1966.

75. Osepchuk, J. M., "Impact of public concerns about low-level EMF-RFR database," in Klauenberg, B. J., Grandolfo, M., and Erwin, D., (eds.)., *Radiofrequency Radiation Standards*, Plenum Press, NY, 425–426, 1995.
76. AFRL, *Digest, Infrared Lasers & Millimeter Waves Workshop: the Links between Microwaves & Laser Optics*, Air Force Research Laboratories, Brooks AFB, TX, January 1997.
77. ANSI C95.1-1982, American National Standard Safety Levels with respect to Human Exposure to Radio Frequency Electromagnetic Fields, 300 kHz to 100 GHz, American National Standards Institute, New York, NY.
78. Wang O., Fujiwara O., "Comparison and evaluation of electromagnetic absorption characteristics in realistic human head models of adult and children for 900 MHz mobile telephones," *IEEE Trans. Microwave Theory Techniques*, vol. 51, pp. 966–971, 2003.
79. Riu P. J., Foster K. R., "Heating of tissue by near-field exposure to a dipole: a model analysis," *IEEE Trans. Biomed Eng.*, vol. 46, pp. 911–917, 1999.
80. Van Leeuwen G. M., Lagendijk J. J., Van Leersum B. J., Zwamborn A. P., Hornsleth S. N., and Kotte A. N., "Calculation of change in brain temperatures due to exposure to a mobile phone," *Phys. Med. Biol.*, vol. 44, pp. 2367–2379, 1999.
81. IEEE C95.2-1999 (Reaffirmed 2005), "IEEE Standard for Radio Frequency Energy and Current Flow Symbols," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2005.
82. IEEE C95.3-2002, "Recommended Practice for Measurements and Computations of Radio Frequency Electromagnetic Fields with Respect to Human Exposure to Such Fields, 100 kHz to 300 GHz," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2002.
83. IEEE C95.6-2002, "IEEE Standard for Safety Levels with Respect to Human Exposure to Electromagnetic Fields, 0 to 3 kHz," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2002.
84. Public Law 90-602, *Radiation Control for Health and Safety Act of 1968*, Approved by Congress, October 18, 1968.
85. Osepchuk, J. M., "Review of microwave oven safety," *J. Microwave Power*, vol. 13, pp. 13–26, 1978.
86. The Internet site for FDA/CDRH is: http://www.fda.gov/cdrh/radhlth/index.html
87. Osepchuk, J. M., "The bluetooth threat to microwave equipment," *Microwave World*, vol. 20(1), pp. 4–5, 1999.
88. FCC, Report and Order FCC 96–326, "*In the Matter of Guidelines for Evaluating the Environmental Effects of Radiofrequency Radiation*," ET Docket No. 93–62, August 1, 1996.
89. IEEE 1528-2003, "Recommended Practice for Determining the Peak Spatial-Average Specific Absorption Rate (SAR) in the Human Head from Wireless Communications Devices: Measurement Techniques," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2003.
90. IEEE 1528a-2005, "IEEE Recommended Practice for Determining the Peak Spatial-Average Specific Absorption Rate (SAR) in the Human Head from Wireless Communications Devices: Measurement Techniques. Amendment 1: CAD File for Human Head Model (SAM Phantom)," Institute of Electrical and Electronics Engineers, Piscataway, NJ, 2005.
91. CEI/IEC 62209-1, "Human Exposure to Radio Frequency Fields from hand-held and Body-mounted Wireless Communication Devices—Human Models, Instrumentation, and Procedures. Part 1: Procedure to Determine the Specific Absorption Rate (SAR) for Hand-held Devices Used in Close Proximity to the Ear (Frequency Range of 300 MHz to 3 GHz)," International Electrotechnical Commission, Geneva, 2005.
92. Dimbylow, P. J. and Mann, S. M., "SAR calculations in an anatomically based realistic model of the head for mobile communication transceivers at 900 MHz and 1.8 GHz," *Phys. Med. Biol.*, vol. 39, pp. 1537–1553, 1994.
93. Okoniewski, M. and Stuchly, M.A., "A study of the handset antenna and human body interaction," *IEEE Trans. Microwave Theory Techniques*, vol. 44(10), pp. 1855–1864, 1996.
94. Olley, P. and Excell, P. S., "Classification of a high resolution voxel image of a human head," *Voxel Phantom Development, Proceedings of an International Workshop held at the National Radiological Protection Board*, Chilton, UK, P. J. Dimbylow, (ed.), 1995.
95. EN50361, "Basic Standard for the Measurement of Specific Absorption Rate Related to Human Exposure to Electromagnetic Fields from Mobile Phones (300 MHz to 3 GHz)," European Committee for Electrotechnical Standardization (CENELEC), Brussels, 2000.

96. Brodeur, P., *The Zapping of America*, William Morrow, NY, 1978.
97. National Research Council, *An Assessment of potential Health Effects from Exposure to PAVE PAWS Low-level Phased Array radiofrequency Energy*, The National Academy Press, Washington, D.C., 2005. Available at Web site www.nap.edu
98. Steneck, N. H., *The Microwave Debate*, MIT Press, Cambridge, MA, 1985.
99. Morgan, M. G., "Prudent avoidance," *Public Utilities Fortnightly*, pp. 26–29, 1992.
100. Park, R., *Voodoo Science*, Oxford University Press, NY, 2000.
101. Foster, K.R., Vecchia, P., and Repacholi, M. H., "Science and the precautionary principle," *Science*, vol. 288, pp. 979–981, May 12, 2000.
102. Commission of the European Communities (European Commission-2000), *Communication on the Precautionary Principle*, Brussels 02 Feb. 2000, See http://www.europa.eu.int/comm/off/cpm/health-consumer/precaution.html. This document refers to ESTO reports at : http://www.jrc.cfapp/reports/list.cfm
103. IEEE ICES (SCC-28), *Response of: Institute of Electrical and Electronic Engineers Standards Coordinating Committee 28 to: Interagency Committee re: Advice on Mitigation of Risks Related to EMF Exposure*, December 1999.
104. Graham, J. D. and Hsia, S., "Europe's precautionary principle: promise and pitfalls," *J. Risk Res.*, vol. 5(4), pp. 37–290, 2002.
105. Osepchuk, J. M., "COMAR after 25 years: still a challenge!," *IEEE Eng. Med. Biol.*, vol. 15(3), pp. 120–125, 1996.
106. Siegel, P. H., "Terahertz technology in biology and medicine," *IEEE Trans. Microwave Theory Techniques*, vol. 52(10), pp. 2438–2447, October 2004.
107. Giri, D. V., *High-Power Electromagnetic Radiation: Non-lethal Weapons and Other Applications*, Harvard University Press, Cambridge, MA, 2004.
108. Beason, D., *The E-Bomb*, Da Capa Press, Perseus Books Group, Cambridge, MA, 2005.

22

Signal Characterization and Modulation Theory

John F. Sevic
Maury MW

Spectral efficiency is of paramount importance when considering the design of virtually all commercial wireless communication systems, whether for voice, video, or data. Spectral efficiency can be measured by the number of users per unit of spectrum or by the number of bits that can be represented per unit of spectrum. In general, wireless service providers are interested in maximizing both the number of users and the number of bits per unit spectrum, which in both cases results in the maximum revenue.

Spectral efficiency can be increased by using several methods including signal polarization, access method, modulation method, and signal coding technique. The first method is commonly adopted in satellite communication systems, where, for example, the uplink and downlink may be right-hand and left-hand circularly polarized, respectively.

Access method refers to how a common resource, such as frequency, is shared among each of the users of the system. Frequency domain multiple access (FDMA) is the basis for all commercial broadcast and most wireless communication systems. With FDMA, each user is allocated a particular section of spectrum that is devoted to that user only. The first-generation cellular phone system in the United States, called advanced mobile phone system (AMPS) has approximately 600 30-kHz channels, each one of which can be utilized for voice and low data-rate communication [1].

Users can also be allocated a certain segment of time, called a slot, leading to time-domain multiple access (TDMA). With TDMA, users share a common frequency and are assigned one, or in some cases more than one slot out of several available slots. In this fashion spectral efficiency is increased by segregating users in time. Second-generation wireless communication systems based on TDMA include the North American Digital Cellular system (NADC) and the Global Standard for Mobile Communications (GSM), which have three slots and eight slots, respectively [2,3].

Code-division multiple access (CDMA) results when each user is assigned a unique code, which is orthogonal to all the other available codes. The Walsh function has been adopted as the orthogonal code set for first-generation CDMA systems. At the receiver, the signal is correlated with a known Walsh function, and the transmitted information thus extracted. The ability of these systems to improve spectral efficiency relies on the development of robust orthogonal functions, since any cross-correlation results in performance degradation. Many wireless systems are hybrids of two access methods. For example, the first CDMA-based wireless system, developed by Qualcomm, is based on FDMA and CDMA [4].

Modulation is the process of impressing an information source on a carrier signal. Three characteristics of a signal can be modulated: amplitude, frequency, and phase. In many types of modulation, two characteristics are modulated simultaneously. Analog modulation results when the relationship between the information source and the modulated signal is continuous. Digital modulation results when the modulated characteristic assumes certain prescribed discrete states.

Signal coding techniques are many and varied, and require detail beyond what can be presented in this introductory chapter. Coding can impact the characteristics of the signal, as will be examined in more detail with CDMA.

The purpose of this chapter is to provide an introduction to the representation and characterization of signals used in contemporary wireless communication systems. The time domain representation of an information-bearing signal determines uniquely what the impact of nonlinear amplification will be, so the study begins with a review of time domain signal analysis techniques. Since the effects of nonlinear amplification are of most interest in the frequency domain, the signal analysis review will also include frequency domain methods. Random process theory is an integral element of digital modulation theory, and will also be covered. Following this, a review of several types of modulations will be given, with both a time domain and frequency domain complex envelope description. The impact of filtering, for spectral efficiency improvement, will be assessed. A probabilistic time domain method of characterizing the envelope of a signal, called the envelope distribution function (EDF), will be introduced. This function is more useful than the peak-to-average ratio in estimating the impact of a PA on a signal. A complete reference section is given at the end of the chapter.

22.1 Complex Envelope Representation of Signals

Using Fourier analysis, any periodic signal can be exactly represented as an infinite summation of harmonic phasors [5]. Most often a signal $x(t)$ is approximated as a finite summation of harmonic phasors, in which case

$$x(t) = \sum_{k=-\infty}^{\infty} \tilde{a}_k e^{j\omega kt} \approx \sum_{k=-Q}^{Q} \tilde{a}_k d^{j\omega kt} \tag{22.1a}$$

where

$$\tilde{a}_k = \frac{1}{T}\int_{t}^{t+T} x(\tau) e^{-j2\pi f\tau} d\tau \tag{22.1b}$$

and Q is chosen sufficiently large to accurately represent the signal under consideration. A physical basis for this approximation is that all systems have an essentially low-pass response.

Parsavel's theorem states that average power is invariant with respect to which domain it is calculated in, and is expressed as

$$\overline{P} = \frac{1}{T}\int_{t}^{t+T} \left|x(\tau)\right|^2 d\tau = \sum_{k=-\text{¢}}^{\infty} \left|\tilde{a}_k\right|^2 \approx \sum_{k=-Q}^{Q} \left|\tilde{a}_k\right|^2 \tag{22.2}$$

In many instances, calculation of average power may be easier to evaluate in either the time domain or the frequency domain representation.

The modulation property illustrates the frequency-shifting nature of time domain multiplication. This is expressed as

$$m(t)\cos(2\pi f_o t) \Rightarrow \frac{1}{2}\left[M(f-f_o)+M(f+f_o)\right] \tag{22.3}$$

where $M(f)$ is the Fourier transform of $m(t)$. Note that both upper and lower sidebands are generated, indicating the presence of a negative frequency component.

In many instances, knowledge of the envelope of a signal is sufficient for its characterization and for assessing the associated impact of a nonlinear PA. An arbitrarily modulated signal can be represented in the time-domain as

$$x(t) = real\left\{m(t)\exp\left[2\pi f_o t + \phi(t)\right]\right\} \tag{22.4}$$

where $m(t)$ and $\phi(t)$ describe the time-varying amplitude and phase of the information signal, respectively, and f_o is the carrier frequency [6]. Note that frequency modulation results by differentiation of the phase modulation. The complex envelope of Equation 22.4 is

$$\tilde{x}(t) = i(t) + jq(t) = m(t)e^{j\phi(t)} \tag{22.5}$$

where $i(t)$ and $q(t)$ are defined as the in-phase and quadrature components of the complex envelope. Although there are other methods available of representing modulated signals, Equation 22.5 is adopted here due to the elegant geometric interpretation afforded by the complex plane representation. The components of Equation 22.5 are calculated from

$$i(t) = real\left\{x(t)e^{-j2\pi f_0 t}\right\} \tag{22.6a}$$

$$q(t) = imag\left\{\hat{x}(t)e^{-j2\pi f_0 t}\right\} \tag{22.6b}$$

where $\hat{x}(t)$ is the Hilbert transform of $x(t)$ [7]. Since statistically independent signals are orthogonal, calculation of the Hilbert transform is often unnecessary, and instead two statistically independent data sources for i, (t) and $q(t)$ can be used. Figure 22.1 shows the spectrum of an arbitrary complex envelope equivalent power density spectrum. Note that the spectrum is not even-symmetric about the y-axis, with the resultant requirement that the time domain signal Equation 22.5 is in general complex.

Linear network analysis using the complex envelope is similar to conventional network analysis. Taking the Fourier transform of Equation 22.5, and denoting the complex transfer function as $\tilde{H}(f)$, we have

$$\tilde{Y}(f) = \tilde{H}(f)\tilde{X}(f) \tag{22.7a}$$

This response can also be calculated directly in the time domain using the convolution integral

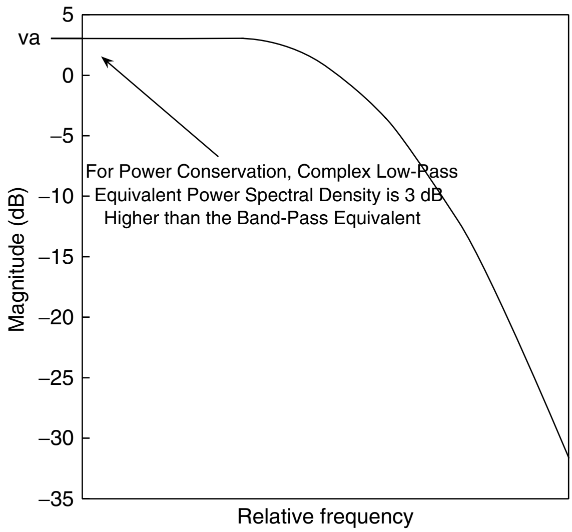


FIGURE 22.1 Power spectral density of an arbitrary complex envelope equivalent signal.

$$\tilde{y}(t) = \int_{-\infty}^{\infty} \tilde{h}(\tau)\tilde{x}(t-\tau)d\tau \tag{22.7b}$$

The equivalent band-pass time domain response is found by

$$y(t) = real\left\{\tilde{v}(t)e^{j2\pi f_0 t}\right\} \tag{22.8}$$

22.2 Representation and Characterization of Random Signals

Consider a random signal $x(t)$ that could describe either a voltage or current. In general the associated *n-dimensional* joint probability density function is required to describe $x(t)$ over n time instants [8]. Since the expected amplitude and power of a signal are most often of interest, the first- and second-moments are sufficient for representation and characterization of $x(t)$. The first- and second-moments of $x(t)$ are

$$\bar{x}(t) = \int_{-\infty}^{\infty} \tau f(\tau)d\tau = \sqrt{DC\ Power} \tag{22.9a}$$

$$\overline{x^2} = \int_{-\infty}^{\infty} \tau^2 f(\tau)d\tau = Total\ Average\ Power \tag{22.9b}$$

where $f(\tau)$ is the associated probability distribution function (pdf) of $x(t)$. The pdf describes the probability of the instantaneous amplitude of $x(t)$ being less than a specified value. This idea is illustrated with a Gaussian pdf in Figure 22.2. Note that the amplitude of this signal is concentrated around its mean value, which is zero.

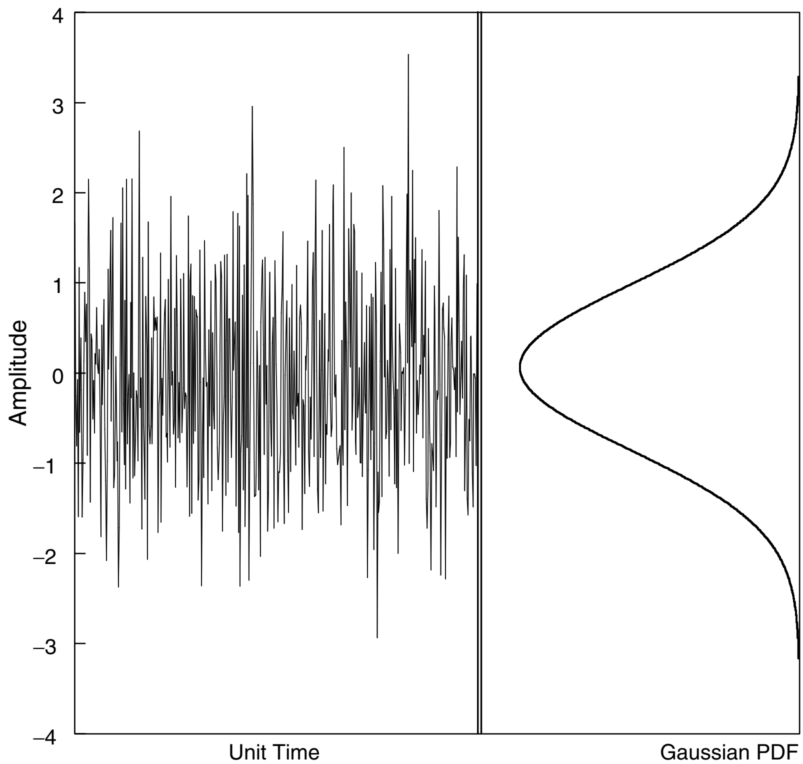


FIGURE 22.2 Illustration of the meaning of random signal having a Gaussian amplitude distribution. In this example the Gaussian pdf represents the probability of the signal amplitude being less than a given value. For example, since the pdf is symmetric about the time axis, this signal has zero mean. Since the noise in this example is uncorrelated, it is defined as white. Similarly, a colored noise process will exhibit correlation.

Since the average value and average power of $x(t)$, as determined by Equation 22.9, are based on ensembles, they are defined as ensemble averages. The average value and average power of a signal can also be calculated with respect to time, which is the method most engineers are familiar with. Signals are said to be ergodic when their ensemble and time averages are the same. All of the signals described in this chapter are ergodic.

Probabilistic characterization of how rapidly a signal changes over time, and hence its spectral distribution, is described by its autocorrelation function

$$R_x(\tau) = \overline{x(t)x(t+\tau)} = \iint_{\infty} \tau_1\tau_2 f(\tau_1, \tau_2)\,d\tau_1 d\tau_2 \tag{22.10}$$

The autocorrelation function evaluated at $\tau = 0$ is the total average power in the signal, as Equation 22.9b shows. Using Equations 22.9a and 22.9b, which represent average amplitude and total average power, respectively, gives the AC power of the signal

$$P_{AC} = R_x(0) - (\bar{x})^2 = \overline{x^2} - (\bar{x})^2 \tag{22.11}$$

Using the Wiener–Khintchine theorem, the spectral distribution of power of $x(t)$ is evaluated by taking the Fourier transform of the autocorrelation function

$$S_x(f) = \int_{-\infty}^{\infty} R_x(\tau) e^{j2\pi f\tau} d\tau \tag{22.12}$$

where $S_x(f)$ is the power spectral density (PSD) of $x(t)$.

In the case of an arbitrary autocorrelation function, $R_x(\tau)$, the PSD is expressed as

$$S_x(f) = S_{pulse}(f) S_{corr}(f) \tag{22.13}$$

where $S_{pulse}(f)$ is the PSD of the pulse function representing $x(t)$ and $S_{corr}(f)$ is the PSD of the autocorrelation function of the data [9]. From this expression, it is clear that the spectral characteristics of a signal are influenced not only by the pulse characteristics, but also any correlation between adjacent pulses.

Let $x(t)$ be the input to a linear system and let $y(t)$ be the associated output. The impulse response, $h(t)$, is described in the frequency domain as $H(f)$. The average amplitude of $y(t)$ is

$$\bar{y}(t) = \tilde{H}(0)\bar{x}(t) \tag{22.14}$$

and the average power of $y(t)$ is

$$\overline{y^2}(t) = R_y(0) = \int_{-\infty}^{\infty} |H(f)|^2 df = \int_{-\infty}^{\infty} |H(f)|^2 S_x(f) df \tag{22.15}$$

The output autocorrelation function is

$$R_y(\tau) = h(\tau) * h(-\tau) * R_x(\tau) \tag{22.16}$$

where * denotes convolution. Taking the Fourier transform gives the output PSD in terms of the input PSD

$$S_y(f) = |H(f)|^2 S_x(f) \tag{22.17}$$

Complex envelope equivalent analysis is done by replacing all variables with the associated complex envelope representation and restricting the integration of Equation 22.15 to positive frequency [6]. Note that this analysis gives an example of how choosing the domain in which to carry out an analysis can greatly simplify the effort involved.

22.3 Modulation Theory

Modulation theory provides a framework for representing and characterizing the time domain and frequency domain characteristics of an information-bearing signal, and the subsequent impact of nonlinear amplification of the signal. The general analysis method to be followed is to describe the signal in the time domain, using the geometric interpretation of Equation 22.5, and then determine resultant signal degradation by characterization in the frequency domain.

Access method, modulation, and coding each directly impact the spectral efficiency of a signal. From Fourier analysis it is also clear, therefore, that the time domain characteristics are also impacted, due to the inverse relationship between the time domain and frequency domain. In other words, a signal that varies rapidly in time, due to access method, modulation, or coding, will be wider in extent in the frequency domain than a slowly varying signal. Note here that signal refers to the envelope of the carrier, and the extent of the signal in the frequency domain refers to the associated spectral description of the envelope.

As Equation 22.5 indicates, modulation can be interpreted geometrically by associating a position in the complex plane with the instantaneous value of the information source. For analog modulation, a continuous range of values is possible; digital modulation allows only certain locations to be occupied. This mapping operation will directly establish the resultant time domain characteristics of the modulation. Many modulations are based on phase since they are relatively impervious to amplitude-related noise disturbances, FM being an example of this. When phase modulation is digital, the characteristics of the digital pulse will influence the signal as well, as Equation 22.13 shows. Since pulses may be rapidly varying, it is expected that some type of filtering will be necessary to give acceptable spectral efficiency. Thus, to describe a modulation, the mapping method and any associated band-limiting filtering must be considered to provide a complete time domain description of the complex envelope.

22.3.1 Analog Modulation

The simplest modulation is analog double sideband suppressed carrier (DSB-SC). When the information source is a sinusoid, DSB-SC is the classical two-tone intermodulation test signal. The complex envelope representation of DSB-SC is

$$\tilde{x}(t) = m(t) \tag{22.18}$$

where $m(t)$ is the information signal. Since the envelope of DSB-SC varies in direct proportion to $m(t)$, it is not a constant envelope modulation. Note also that, from a geometric interpretation using Equation 22.5, DSB-SC requires only one dimension, meaning there is no phase modulation.

All of the first-generation cellular systems, such as AMPS and ETACS, are based on FM, with a complex envelope representation of

$$\tilde{x}(t) = \exp\left[jk_f \int_{-\infty}^{t} m(\tau)\,d\tau\right] \tag{22.19}$$

where k_f is the frequency-deviation constant and $m(t)$ is the information signal. Since the magnitude of the complex exponential is unity, FM is a constant envelope modulation. A geometric interpretation of FM shows that it would be a unit circle, with the speed in movement about this circle proportional to the instantaneous value of the information signal. Carson's rule shows that the spectral efficiency of FM is less than DSB-SC [6]. In general, constant envelope modulation is not as spectrally efficient as modulation that exhibits a time-varying envelope.

22.3.2 Discontinuous Phase-Shift Keying

Virtually all second- and third-generation wireless communication systems are based on digital modulation, with digital phase modulation being the most common. Digital phase modulation is commonly referred to as phase-shift keying (PSK). PSK consists of two DSB-SC signals in quadrature, and can be represented using Equation 22.5 with in-phase and quadrature components each generated from a digital data source.

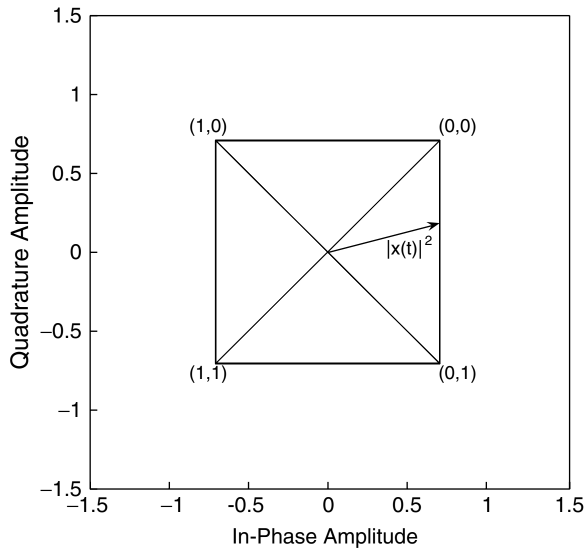


FIGURE 22.3 Complex plane representation of unfiltered quadrature modulation (QPSK). Each corner of the constellation diagram represents two data bits. Depending on the data sequence, each transition goes through a phase change of either ± 90° or ± 180°, which always causes a step change in the signal envelope.

Each data source is usually generated by multiplexing a serial data stream, which, if necessary, is already coded, such as with CDMA. Quadrature digital modulation is represented in the complex plane as shown in Figure 22.3, where the horizontal axis represents the real part of Equation 22.5 and the vertical axis represents the imaginary part of Equation 22.5. The trajectory is the instantaneous envelope of the signal and for the constellation given, there are four unique phases, with each phase representing a unique combination of two bits. Each combination of bits is defined as a symbol. Some third-generation wireless systems have adopted PSK with eight unique locations, giving 8-PSK, with the result that each location now represents a unique combination of three bits. Note also from Figure 22.3 that PSK modulation is constant envelope.

The complex envelope representation of quadrature digital modulation is

$$\tilde{x}(kT_b) = i(kT_b) + jq(kT_b) \tag{22.20}$$

where

$$i(kT_b) = \sum_{k=-\infty}^{\infty} a(kT_b) f(t - kT_b - \tau_f) \tag{22.21a}$$

$$q(kT_b) = \sum_{k=-\infty}^{\infty} b(kT_b) g(t - kT_b - \tau_g) \tag{22.21b}$$

and $a(kT_b)$ and $a(kT_b)$ represent unit-amplitude data sources, f and g are pulse functions, T_b is the bit rate, and τ_f and τ_g are arbitrary phase offsets. The functions f and g are usually rectangular pulse streams with zero mean and zero correlation between pulses. In this case, the PSD of Equations 22.21a and (22.21b) is given as

$$S_i(f) = T_b \operatorname{sinc}^2(\pi T_b f) \tag{22.21c}$$

$$S_q(f) = T_b \operatorname{sinc}^2(\pi T_b f) \tag{22.21d}$$

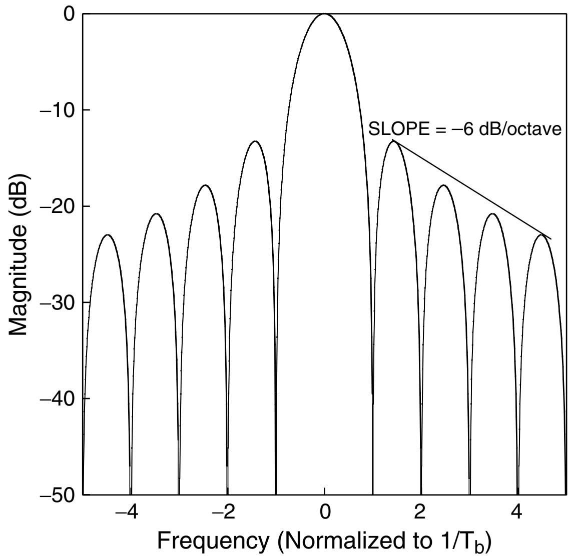


FIGURE 22.4

In Figure 22.4 it is shown that Equations 22.21c and 22.21d exhibit a –6 dB/octave roll off. A communication system based on rectangular pulses would therefore exhibit very poor spectral efficiency due to the relatively large energy present in the spectrum in the neighboring sidelobes. To resolve this problem, band-limiting filtering is used. Although it is possible to use arbitrary band-limiting filters, such as the Chebyshev low-pass response, the Nyquist response is often adopted since it exhibits certain characteristics amenable to demodulation of the signal at the receiver [10–12].

A significant consequence of band-limiting a discontinuous signal, such as that exhibited by PSK with rectangular pulses, is that envelope variations will be introduced. Thus, although PSK modulation is constant envelope, the necessity of band-limiting induces envelope variations, which gives a band-limited PSK signal properties of both phase modulation and amplitude modulation. The degree of amplitude variation is directly proportional to the degree of band limiting. Figure 22.5 illustrates the impact of band limiting on the QPSK constellation of Figure 22.3, where the amplitude variations are apparent. Band limiting will necessarily require a reduction in the efficiency of a PA, in order to support the peak excursions of the envelope.

The Nyquist filter response exhibits simultaneous band limiting and zero inter-symbol interference (ISI). Zero ISI is desirable, though not necessary, for a digital communication system to give acceptable performance. To understand the impact of inter-symbol interference, consider the convolution integral Equation 22.7b, where it is seen that convolution is essentially a summation operation that accounts for the past state of a signal or system at the present time. Use of an arbitrary filter response will result in interference of the present state of a signal due to its past values, leading to the potential for an error in the actual value of the signal at the present time. The impulse response of a Nyquist filter has zero crossings at multiples of the symbol rate, and thus does not cause inter-symbol interference [11,12].

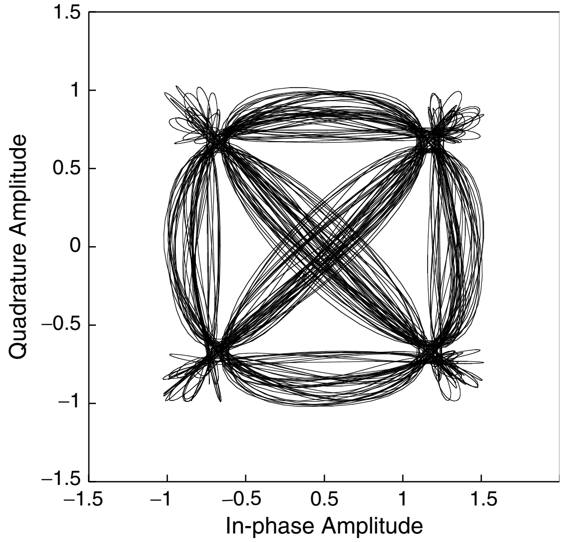


FIGURE 22.5 Constellation diagram of band-limited QPSK. The transient response of the band-limiting filter increases the peak-average ratio of the signal. The more severe the band limiting, yielding increased spectral efficiency, the higher the peak-to-average ratio becomes.

The most common form of Nyquist filter currently adopted is the raised cosine response. The raised-cosine response is expressed in the frequency domain as

$$H(f) = \begin{cases} T_s, & |f| \le \dfrac{1}{2T_s} - \alpha \\ T_s \cos^2\left[\dfrac{\pi}{4\alpha}\left(|f| - \dfrac{1}{2T_s} + \alpha\right)\right], & \dfrac{1}{2T_s} - \alpha \le |f| \le \dfrac{1}{2T_s} + \alpha \\ 0, & |f| > \dfrac{1}{2T_s} + \alpha \end{cases} \tag{22.22}$$

where T_s is the symbol rate in seconds and α is the excess-bandwidth factor [6,13]. By adjusting α, the spectral efficiency of a digitally modulated signal can be controlled, while maintaining zero ISI. Figure 22.6 shows the frequency response of Equation 22.22 for various values of α. At the Nyquist limit of $\alpha = 0$, all energy above half of the main lobe is removed. This results in maximum spectral efficiency but the largest time domain peak-to-average ratio.

To maximize receiver signal-to-noise ratio (SNR), many wireless communication systems split the filter response equally between the transmitter and receiver [9,10]. The resultant response is called the square-root raised-cosine response, which in the time domain is expressed as

$$h(t) = 4\alpha \frac{\cos\left[(1+\alpha)\dfrac{\pi}{T_s}t\right] + \dfrac{\sin\left[(1-\alpha)\dfrac{\pi}{T_s}t\right]}{\dfrac{4\alpha}{T_s}t}}{\left(\pi T_s^{0.5}\right)\left[\left(\dfrac{4\alpha}{T_s}t\right)^2 - 1\right]} \tag{22.23}$$

where the variables are defined as above in Equation 22.22 [11].

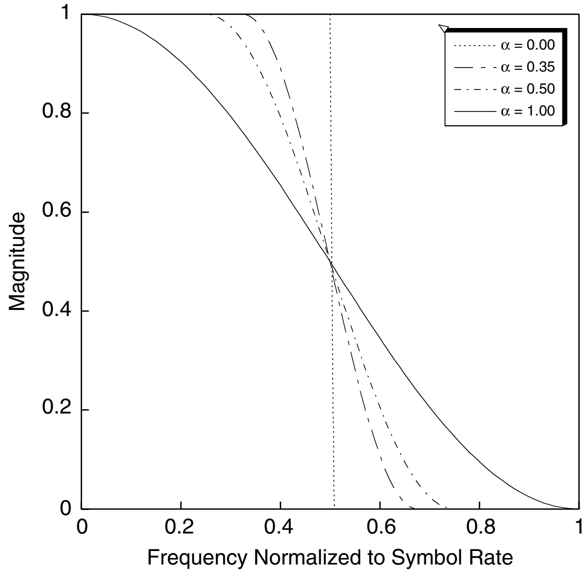


FIGURE 22.6 Raised-cosine filter response for several excess-bandwidth factors. The frequency axis is normalized to the symbol rate. NADC uses an $\alpha = 0.35$.

Many digital wireless standards specify the symbol rate and excess-bandwidth factor to specify the root-raised cosine filter. This can lead to ambiguity in specifying the filter response insofar as the length of the impulse response is not specified. An impulse response that is too short will exhibit unacceptable band-limiting performance and can lead to incorrect characterization of spectral regrowth. Alternatively, some standards specify a band-limiting impulse response in the form of a digital filter to ensure that the appropriate frequency domain response is used [4,14].

Several wireless communication systems are based on discontinuous PSK modulation. These include the North American Digital Cellular standard (NADC), the Personal Digital Cellular System (PDC), the Personal Handyphone System (PHS), and cdmaOne [2,4,15]. The first three of these systems have adopted $\pi/4$-DQPSK modulation with a root-raised cosine filter response. The cdmaOne system has adopted QPSK for the forward-link and offset QPSK (O-QPSK) for the reverse link, in both cases using a Chebyshev filter response.

$\pi/4$-DQPSK, a derivative of QPSK, differentially encodes the symbols, and adds a $\pi/4$ rotation at each symbol instant. This results in an envelope that avoids the origin, thus decreasing the peak-average of the signal. This is done presumably to reduce the PA back-off requirement, but as recent work has illustrated, this is not achieved in practice due to the original assumptions made [16]. Figure 22.7 shows the constellation diagram for $\pi/4$-DQPSK with $\alpha = 0.35$, corresponding to the NADC standard. Note that at each of the constellation points, the trajectories do not overlap; this is inter-symbol interference. At the receiver, this signal will be convolved with the remaining half of the raised-cosine filter, leading to the desired response given by Equation 22.22. PDC and PHS are also based on $\pi/4$-DQPSK, with an excess bandwidth factor of $\alpha = 0.50$. Their constellation diagrams are similar to the NADC diagram shown in Figure 22.7.

Figure 22.8 shows the constellation diagram for reverse-link cdmaOne, which is based on O-QPSK. Like $\pi/4$-DQPSK, O-QPSK was developed to reduce the peak-to-average ratio of the signal. It has been shown that this approach, while reducing the peak-to-average ratio, does not necessarily reduce the PA linearity requirements [16]. Note also in Figure 22.8 the significant ISI, due to the fact that a Nyquist filter was not used. The forward link for cdmaOne uses QPSK modulation, which is shown in Figure 22.5. As with the reverse link, the forward link uses a Chebyshev response specified by the standard.

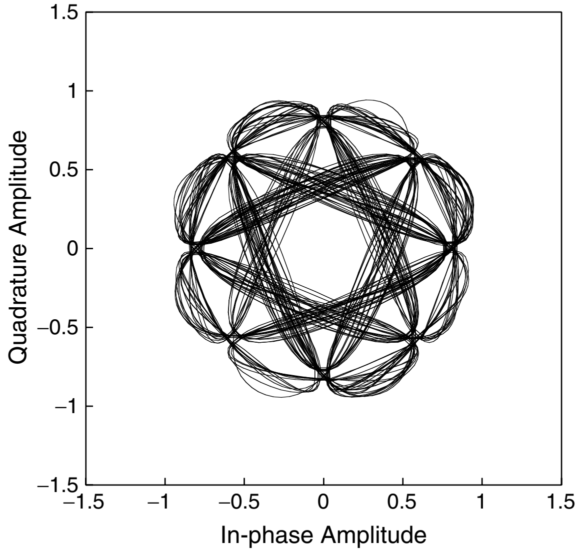


FIGURE 22.7 Constellation diagram for $\pi/4$-DQPSK with $\alpha = 0.35$. This modulation is used for NADC, PDC, and PHS.

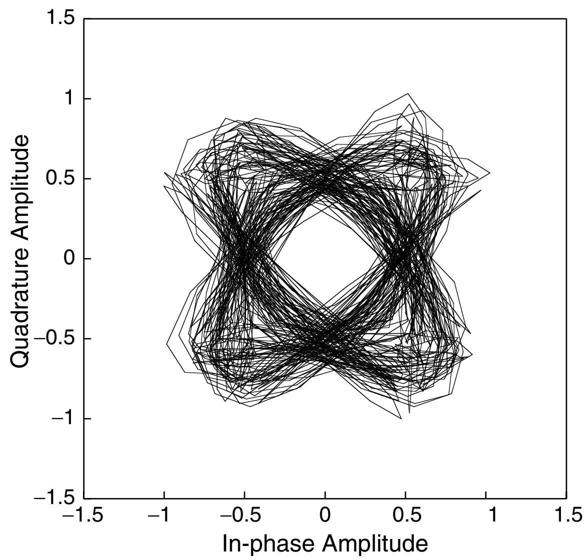


FIGURE 22.8 Constellation diagram for O-QPSK using the filter response specified by the IS-95 standard for cdmaOne.

To increase spectral efficiency, it is possible to add four additional phase locations to QPSK, called 8-PSK, with each location corresponding to a unique combination of three bits. Figure 22.9 shows the constellation diagram for the EDGE standard, which is based on 8-PSK and a specially developed filter that exhibits limited response in both the time domain and frequency domain [17]. EDGE was developed to increase the system capacity of the current GSM system while minimizing the requirements for linearity. In contrast to GSM, which is constant envelope, the EDGE systems provides increased spectral efficiency at the expense of envelope variations. However, EDGE intentionally minimizes the resultant amplitude modulation to ensure that existing GSM amplification systems, optimized for constant envelope modulation, could still be used.

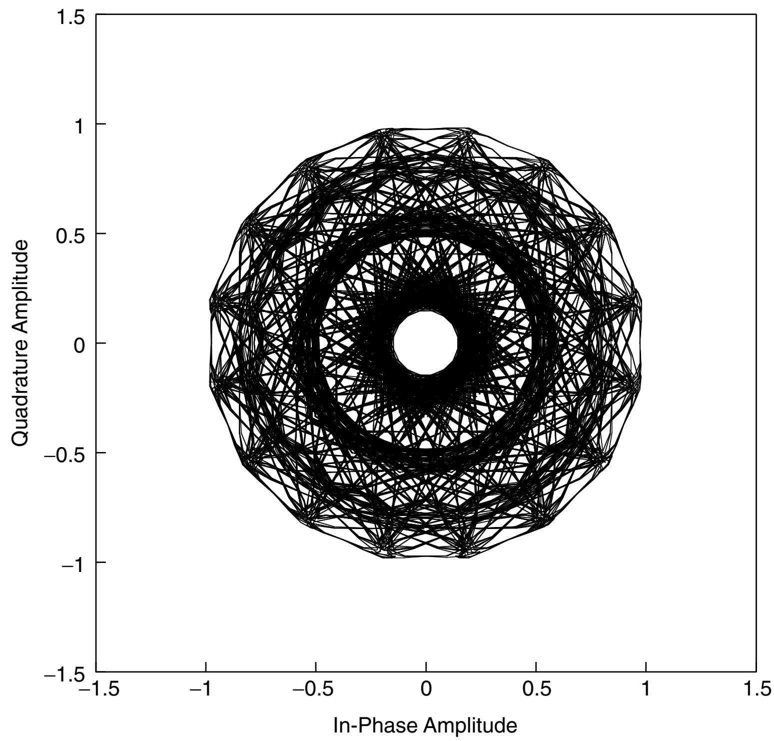


FIGURE 22.9 Constellation diagram for EDGE.

22.3.3 Continuous Phase-Shift Keying

With the exception of EDGE, each of the PSK modulations considered in the previous section exhibited significant out-of-band energy due to the discontinuous transitions. To reduce the out-of-band energy, band limiting was required, which resulted in envelope variations. An alternative modulation process is continuous-phase modulation (CPM), which requires that the phase transitions from one symbol to the next be continuous [9,13]. This results in a signal that is intrinsically band limited while maintaining a constant envelope. The price paid is a potential reduction in spectral efficiency. Two types of CPM will be examined in this section: minimum-shift keying (MSK) and Gaussian MSK (GMSK).

The complex envelope representation of MSK is

$$x(t) = \cos\left[a\left(kT_b \frac{\pi t}{2T_b}\right)\right] + j\sin\left[b(kT_b)\frac{\pi t}{2T_b}\right] \tag{22.24}$$

$a(kT_b)$ and $b(kT_b)$ represent unit-amplitude data sources as defined in Equation 22.21. This modulation is identical to O-QSPK using a half-wave sinusoidal pulse function instead of a rectangular pulse function. Taking the magnitude squared of each term of Equation 22.24 shows that it is indeed a constant envelope.

GMSK uses Gaussian pulse shaping, and has a complex envelope representation of

$$x(t) = \exp\left[jk_f \int_{-\infty}^{t} i(\tau)d\tau\right] = \exp\left[jk_f \int_{-\infty}^{t} \sum_{k=-\infty}^{\infty} a(kT_b)f(\tau - kT_b)d\tau\right] \tag{22.25}$$

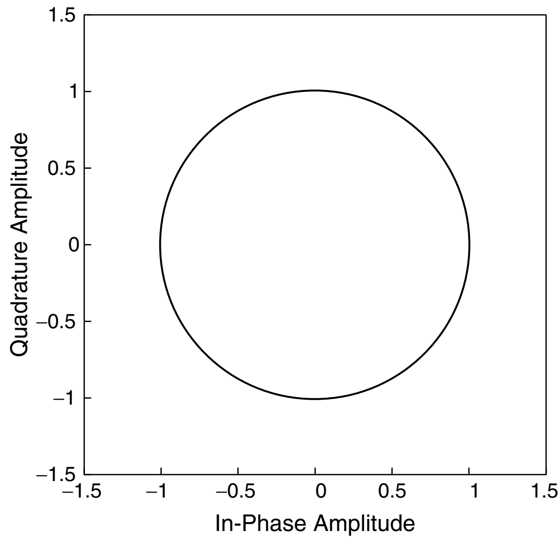


FIGURE 22.10 Constellation diagram for GMSK, which is used in the GSM wireless system.

where $f(\tau - kT_b)$ is a Gaussian pulse function and $a(kT_b)$ is a unit-amplitude data source as defined in Equation 22.21. Since the Fourier transform of a Gaussian pulse in the time domain is a Gaussian pulse, it is seen that this modulation will exhibit intrinsic band limiting, in contrast to PSK. In Figure 22.10 the GMSK constellation diagram is illustrated, where it is seen that the envelope is constant. The information is contained in how rapidly the phase function moves from one location on the circle to another, in a fashion similar to FM. GMSK is used in the Global Standard for Mobile Communications (GSM) wireless system [3].

22.4 Probabilistic Envelope Characterization

The complex trajectory of a signal, determined by the modulation technique, band limiting, and signal coding used, is a time parametric representation of the instantaneous envelope of the signal. As such, the duration of a particular envelope level in combination with the transfer characteristics of the power amplifier (PA) will establish the resultant instantaneous saturation level. If the average spectral regrowth exhibited by a PA is considered a summation of many instantaneous saturation events, it follows that the more often an envelope induces saturation, the higher the average spectral regrowth will be. It is for this reason that the peak-to-average ratio, though widely used, is ill suited for estimating and comparing the linearity requirements of a PA [16].

This section introduces a method for probabilistically evaluating the time domain characteristics of a an arbitrary signal to establish its associated peak-to-average ratio and instantaneous envelope power distribution. The envelope distribution function (EDF) is introduced to estimate the peak power capability required of the PA and compare different signals. The EDF for many of the wireless systems presently in use are examined.

22.4.1 The Envelope Distribution Function

Let $\tilde{p}(t)$ be the instantaneous power of a signal $\tilde{x}(t)$, with an associated instantaneous power probability distribution function $\varphi(\tilde{p})$. The probability of the instantaneous power exceeding the mean signal power is

$$\Pr\left[\text{instantaneous power} > \text{mean power}\right] = \Psi\left(\tilde{p}\right) = 1 - \int_{E[\tilde{p}]}^{\infty} \varphi\left(\tilde{p}\right) d\tilde{p} \tag{22.26}$$

where $E[\tilde{p}]$ is the average power of $\tilde{p}(t)$. This function is defined as the envelope distribution function (EDF). In practice, the EDF is evaluated numerically, although it is possible to generate closed-form expressions. A specified probability of the EDF, typically 10^{-6}, is referred to as the peak-to-average ratio, σ. This is defined as

$$\sigma = \frac{\text{EDF @ } 10^{-6}}{E\left[\tilde{p}(t)\right]} \tag{22.27}$$

A gradual roll off of the EDF indicates the instantaneous power is less likely to be near the mean power of the signal. This characteristic implies enhanced linearity performance to minimize increased distortion associated with the relative increase in the instantaneous clipping of the signal. Alternatively, for a given linearity requirement, the average power of the signal must necessarily decrease, resulting in lower efficiency. The amount that the average power must be reduced is referred to as output back off, and is usually specified in dB normalized to the 1 dB compression point of the PA under single-tone excitation. Finally, note that the EDF only has meaning for signals with a time-varying envelope.

22.4.2 The EDF for Various Wireless Systems

In this section, Equation 22.26 is used to examine the EDF for several of the wireless systems described earlier. NADC, PDC, and EDGE are considered first. The EDF for CDMA-FL and CDMA-RL using only a pilot tone is considered next. The impact of traffic channel, synchronization channel, and paging channels on the CDMA-FL EDF is then illustrated, where it will be seen that a significant increase in the peak-to-average ratio results, making this signal difficult to amplify. For comparison purposes, the EDF for a two-tone signal and complex Gaussian noise will also be shown. The EDF for each of these signals is shown in Figures 22.11 through 22.16.

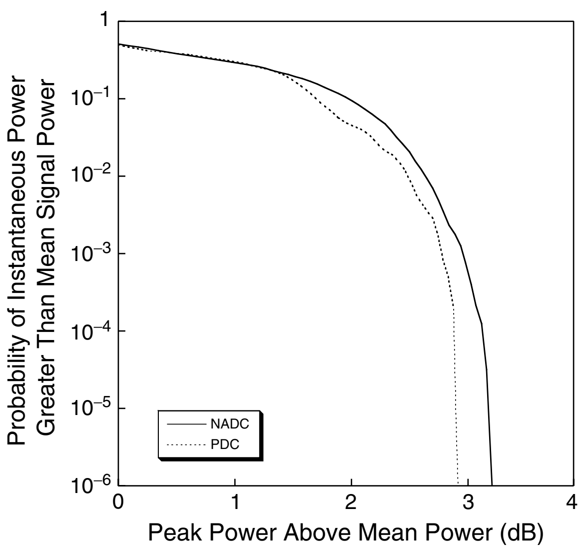


FIGURE 22.11 The EDF for NADC and PDC. NADC uses $\pi/4$-DQPSK with $\alpha = 0.35$ and PDC uses $\pi/4$-DQPSK with $\alpha = 0.50$. Note PDC has a slightly lower peak-to-average ratio due to the more gradual roll off of the filter response.

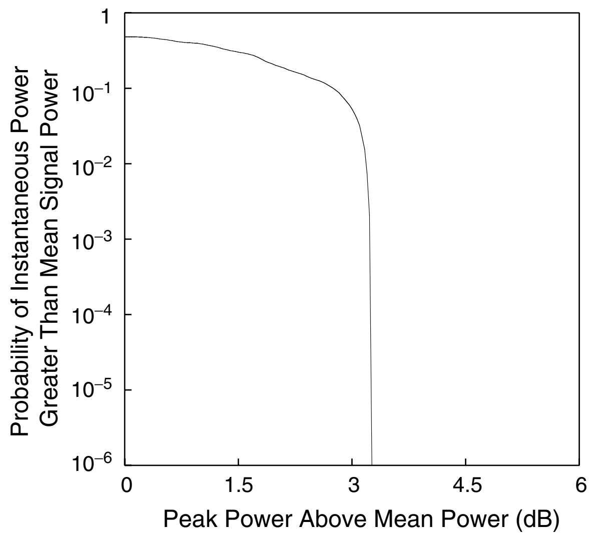


FIGURE 22.12 The EDF for EDGE (with all time slots active). Compare to Figure 22.15, where the EDF for a two-tone signal is shown.

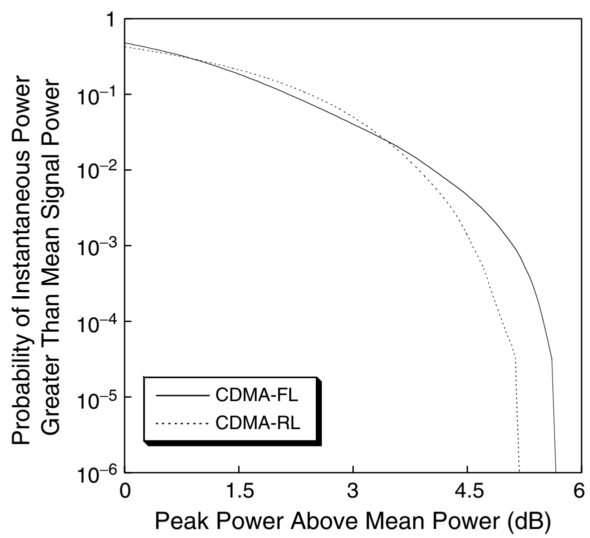


FIGURE 22.13 The EDF for forward-link CDMA and reverse-link CDMA with pilot tone only. Note that although CDMA-RL has a lower peak-to-average ratio, in areas of high probability of occurrence it has a higher peak power and results in higher spectral regrowth than CDMA-FL. This comparison clearly illustrates the advantages of using the EDF over the peak-to-average ratio.

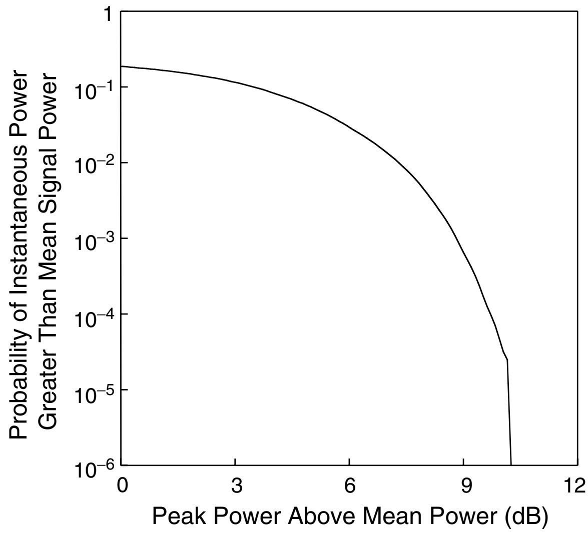


FIGURE 22.14 The EDF for forward-link CDMA with six traffic channels active and synchronization, paging, and pilot also active. Compare to Figure 22.13 and observe the significant increase in peak-to-average ratio.

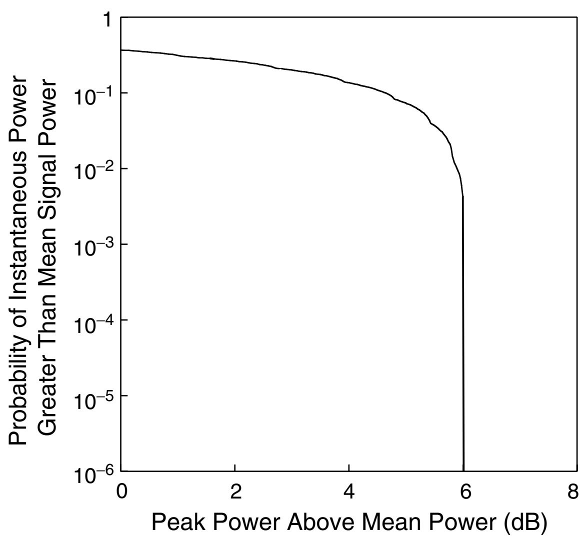


FIGURE 22.15 The EDF for a two-tone signal. Note that like the EDGE signal, the two-tone signal exhibits a gradual roll off, leading to increased distortion with respect to a similar signal with the same peak-to-average ratio but a faster roll off, such as CDMA-FL.

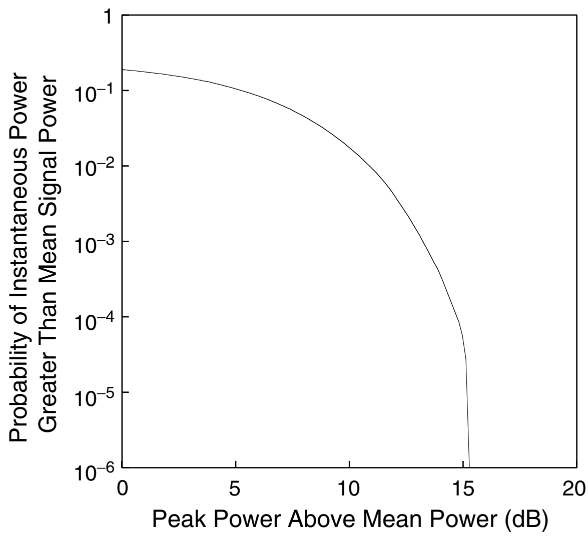


FIGURE 22.16 The EDF for a complex white Gaussian noise signal (commonly used for noise-to-power ratio characterization). The theoretical peak-to-average of this signal is infinite, but in practice is approximately 10 dB to 15 dB, depending on the length the sequence used.

22.5 Summary

Contemporary microwave circuit design requires a basic understanding of digital modulation theory in order to meet the needs of a customer who ultimately speaks in terms of communication theory. This chapter was intended to provide a brief overview of the signal analysis tools necessary for the microwave engineer to understand digital modulation and how it impacts the design and characterization of microwave circuits used in contemporary wireless communication systems.

Complex envelope analysis was introduced as a means to describe arbitrarily modulated signals, leading to a geometric interpretation of modulation. The necessity and subsequent implications of band-limiting PSK signals were discussed. As an alternative to PSK, CPM was also introduced.

Signal analysis methods are often used to simplify the design process. Although the peak-to-average ratio of a signal is widely used to estimate of the linearity requirements of a PA, it was shown that this metric is ill suited in general for this purpose due to the random distribution of instantaneous power of a signal. The EDF was introduced as means to compare various signals and to provide a more accurate estimate of the required linearity performance of a PA.

References

1. W. C. Y. Lee, *Mobile Cellular Telecommunications Systems*, New York: McGraw-Hill, 1989.
2. EIA/TIA Interim Standard, Recommended Minimum Performance Standards for 800 MHz Dual-Mode Mobile Stations, EIA/TIA/IS-55, 1991.
3. M. R. L. Hodges, The GSM Radio Interface, *British Telecom. Tech. Journal,* 8, 31–43, January 1990.
4. EIA/TIA Interim Standard, *Mobile Station—Base Station Compatibility Standard for Dual-Mode Wideband Spread Spectrum Cellular System,* TIA/EIA/IS-95, 1993
5. R. N. Bracewell, *The Fourier Transform and Its Application*, New York: McGraw-Hill, 1978.
6. L. W. Couch, *Digital and Analog Communications*, 4th ed., New York: Macmillan, 1994.

7. M. Jeruchim, P. Balaban, and K. Shanmugan, *Simulation of Communication Systems*, New York: Plenum Press, 1992.
8. A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed., New York: McGraw-Hill, 1991.
9. M. K. Simon, S. M. Hinedi, and W. C. Lindsey, *Digital Communication Techniques: Signal Design and Detection*, New York: Prentice-Hall, 1994.
10. M. Schwartz, *Information Transmission, Modulation, and Noise*, 3rd ed., New York: McGraw-Hill, 1980.
11. W. R. Bennett and J. R. Davey, *Data Transmission*, New York: McGraw-Hill, 1965.
12. H. Nyquist, Certain topics in telegraph transmission theory, *Trans. AIEE*, 47, 117–141, February 1928.
13. J. G. Proakis, *Digital Communications*, 2nd ed., New York: McGraw-Hill, 1992.
14. R. Schaefer and A. Openheim, *Discrete Time Signal Processing*, New York: McGraw-Hill, 1992.
15. Research and Development Center for Radio Systems (RCR), *Digital Cellular Communication Systems,* RCR STD-27, April 1991.
16. J. F. Sevic and M. B. Steer, On the Significance of Envelope Peak-to-Average Ratio for Estimating the Spectral Regrowth Characteristics of an RF/Microwave Power Amplifier, *IEEE Transactions on Microwave Theory and Techniques*, in press.
17. ETSI Standard GSM 11.21: Digital Cellular Telecommunications System BSS Radio Aspects, ETSI, 1999.

23 Productivity Initiatives

Mike Golio
HVVi Semiconductor

23.1 Introduction

Organizational productivity initiatives have become a significant part of the engineering design system utilized by the radio frequency (RF)/microwave industry. These abundant initiatives have modified and molded the microwave engineer's methodologies and work habits in the past 15–20 years. The popularity of these initiatives is so pervasive among companies and engineering organizations that an entire industry of consultant trainers has emerged to help launch and implement successful programs to improve organizational efficiency and performance. Popular initiatives include

- Design for Manufacturability
- Design to Cost
- Total Customer Satisfaction
- Six Sigma
- Just-In-Time Delivery
- Total Cycle Time Reduction
- Total Quality Management
- Lean Electronics
- Best Place to Work
- Supply Chain Operation Reference

All of the aforementioned initiatives are designed to improve productivity—and therefore profit. Some initiatives focus primarily on the use of new tools and metrics (such as Six Sigma) to improve manufacturing productivity. Just-In-Time initiatives focus on inventory reduction and procurement efficiency. Initiatives like Lean Electronics represent a collection of productivity techniques coupled by an underlying theme of efficiency. Finally, some initiatives focus primarily on direct alteration of the attitudes and perceptions of the workforce toward their customers and coworkers (such as Total Customer Satisfaction and Best Place to Work). To some extent, all major productivity initiatives are culture-changing exercises.

Although it is not possible within a single article to discuss (or even list) all of the specific productivity initiatives that have invaded the workplace, it is possible to discuss some of the fundamental engineering principles and tools that lead to efficient engineering practice.

23.2 Customizing Initiatives for the Organization

Improved productivity and profit are undeniably the most compelling reasons for a company to undertake the task of changing its culture. There are many reasons, however, why the successful implementation of a productivity initiative must be tailored to a specific organization. Implementation of any successful initiative must account for (1) the market and market drivers for the organization undergoing change and (2) the existing culture of that organization.

Figure 23.1 illustrates how particular microwave/RF markets affect the primary focus and development time constants of the technology development process. The shift of technical focus is dramatic as technologies move from a proof-of-concept to a high volume manufacturing stage. A similarly dramatic shift in the acceptable development timeframes is associated with this maturity process.

RF and microwave technology typically travels a maturation path that starts in a university, industry, or government R&D lab with a feasibility demonstration. These proof-of-concept efforts sometimes take decades or even lifetimes to achieve. This type of work is focused primarily on the development of a one-time proof-of-concept demonstration. Although the advocates of the new technology may have visions of

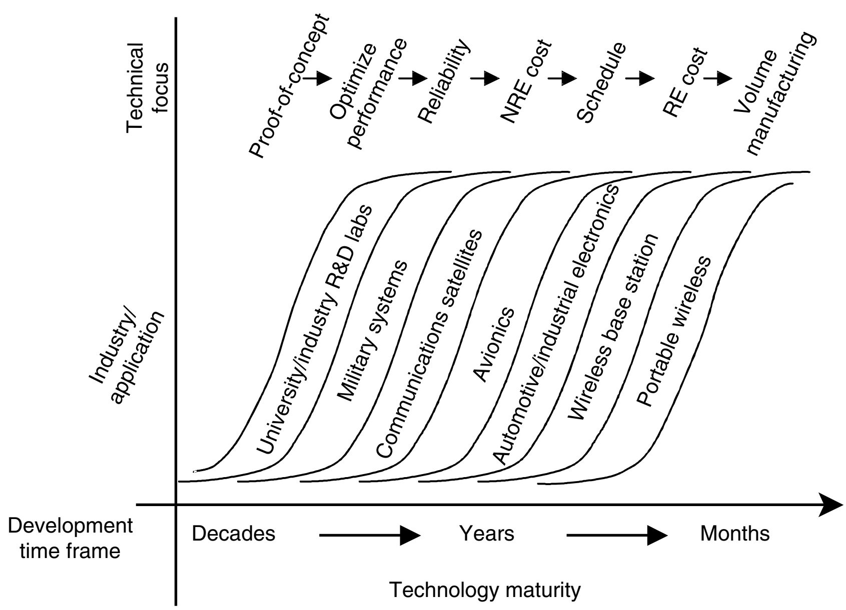


FIGURE 23.1 Illustration of a typical maturation path for emerging RF technologies. At each phase of the maturity process, technical focus and development timeframes are radically different. Productivity initiatives should account for these differences if they are to be implemented successfully.

high volume manufacturing at a future date, the initial goal is simply to prove that the idea has potential technical merit.

Promising technologies with demonstrated proof-of-concept merit typically migrate into low volume, high performance applications. Often government/military needs drive these early applications. In this stage of maturity, the research and development is likely to be focused on optimization of the performance and on reliability issues. Development timeframes often run for several years.

As success is reached at each stage of technology development, new applications emerge with different technical foci and shorter acceptable development times. Communications satellites and avionics are example applications that form the center of the technology maturation chart of Figure 23.1. Manufacturing volumes for these applications are still low enough that the nonrecurring engineering (NRE) costs usually dominate the technical focus of the development process, which can still involve timeframes of many years.

Applications such as automotive, industrial, and wireless base station electronics represent the next stages of typical technology maturation. These applications often support high volumes that are enough to bring increased focus onto development schedule duration and recurring costs (RE).

Portable wireless products represent an application that is dominated by manufacturability issues and reduced cycle times. Product development cycles are typically measured in months while cost, yield, and factory throughput dominate the focus of the technical workforce.

A successful productivity initiative needs to address the primary technical focus and development time constants of the industry where it is being applied. From the previous discussion, it is clear that any profit-driven productivity initiative needs to account for the unique constraints of the organization's target market. Initiatives with primary focus on statistical process control and high manufacturing yields may be critical to an organization working in the Portable Wireless industry, but have very little applicability to a R&D lab with a charter to perform long-term, high-risk research of new technologies.

Even within a given market segment, application of productivity initiatives must be customized to specific organizations. As obvious as this statement is to most engineers, it is often ignored by executives anxious to make their mark in a company or by consultants armed with a presentation package. Different countries, regions of a country, companies, and even organizations within a large company, all of these have different roles, different cultures, and different levels of productivity awareness and maturity. Since these initiatives are culture-changing processes, it is important that the existing culture is understood so that appropriate steps are taken to mature the culture toward the ultimate goal. An organization that has never before utilized any quality or productivity metrics may require a remedial approach in order to make progress. In contrast, an organization with lengthy experience in a wide range of productivity practices may not benefit from anything less than a fully integrated, customer-focused, quality manufacturing system.

23.3 Productivity and Marketing

For most product development applications, the engineering process begins with marketing. Productivity initiatives that focus on customer awareness and customer relationships (e.g., Total Customer Satisfaction) should address some of the marketing issues that concern engineers. Marketing establishes the groundwork for Design for Manufacturability efforts.

Prior to any significant engineering planning, it is important to identify the required sales volumes, cost targets, and acceptable development schedule (i.e., required time-to-market).

As an example, assume that an 18 GHz radio product is scheduled to go into production in 18 months with a cost target of $100 per unit at a quantity of 100,000 units per month. The targeted time-to-market dictates that little new technology will be developed. In only 18 months, there will barely be time to identify a design, find components, assemble and align a working breadboard, and develop a production manufacturing and test plan that is consistent with an existing production facility. The limit placed on the amount of technology development that is feasible in such a project extends not only to the radio components, but also to the manufacturing and test technologies needed in production. An analysis of the quantity requirements (assuming two shifts working five days a week) leads to the result that the radio product must be manufactured at the rate of approximately one unit every 12 s. Clearly, little tuning or manual alignment will be acceptable. The quantity and cost targets dictate an automated manufacturing and test environment. Locating off-the-shelf parts available in the quantities required will be critical. Assuming a reasonable product life cycle, the RE cost of manufacturing dominates this engineering effort. NRE cost of development is likely to be insignificant as compared to the goal of achieving the required development cycle time.

In contrast, the entire engineering approach changes if a total of 4 similar radios are to be produced for an R&D demonstration as part of a 3-year research project. Optimum technical performance may be the primary focus of the effort. A premium is likely to be paid (or even required) for utilizing new technologies, new architectures, or new methodologies. Recurring manufacturing cost has little meaning while NRE cost is the determining factor in the profitability of the program. Delays in schedule of months or years may even be acceptable provided the innovation is great enough.

When optimizing for profit, NRE costs, RE costs, performance, and schedule are all part of the formula. The engineering process must compare the options to the marketing requirements. In consumer applications, where time-to-market is critical, a rapidly developed, unoptimized design may provide enough performance at a low enough cost to succeed. The critical questions for these cases may be "How long will it take to achieve the minimum performance?" and "How long will it take to achieve the lowest cost?" In general, speed and technology insertions are inversely related.

23.4 Planning and Scheduling

Although planning and scheduling are not explicitly addressed in many productivity initiatives, they are the foundation for all engineering efforts. Without good planning, other productivity efforts can produce only modest success at best. Planning and scheduling are also required for the development of important productivity metrics.

Planning and scheduling are the processes of deciding in advance what work is required, and when and where the work will be performed. The scheduling process seeks estimates of the time to do the work, identification of the organizations or people who will perform the work, and identification of the required resources.

There are a number of scheduling tools, methodologies, and representations that can be used to perform the planning and scheduling functions. Gantt systems are perhaps the most common scheduling tools used in engineering. Regardless of the methodology, the baseline schedule represents the yardstick against which actual work is measured. If improvements in engineering productivity are to be quantified, it is critical that planning and scheduling be done as accurately as possible.

Table 23.1 lists several desirable characteristics for a good engineering planning and scheduling process.

TABLE 23.1 Desirable Characteristics of a Good Engineering Planning and Scheduling Process

Characteristic of Schedule and Planning Process	Comments
Formal process	Involve representation from all organizations that will be required to release the product
Simple process	Should be simple enough to encourage use by many contributors
Comprises understandable work tasks	Work should be broken down into tasks small enough to establish estimates of the level of effort
Includes meaningful objectives	Schedule must identify meaningful measurable milestones that mark the end of each work task

The planning process should be a formal, concise process that involves representation from all organizations that will contribute directly to the final release of the product. This will include not only project managers and engineers, but may also require representation from manufacturing, systems engineering, quality assurance, reliability, marketing, contracts, and sales.

Simplicity is another valued characteristic of a good planning and scheduling process. The scheduling system needs to be used and updated by a large number of people. Complexity or confusion should not discourage use of the system. Unneeded complexity will also add cost to the process. The scheduling system should provide the required information at the lowest cost.

All schedules start from the documentation of clearly identified objectives. A well-defined project plan expressed in a work breakdown structure (WBS) can accomplish this. The WBS should be composed of work tasks that end with the completion of a meaningful milestone. The work tasks that comprise the WBS must be developed with enough understanding to establish accurate estimates of the time and effort required to complete the task.

Establishing meaningful milestones is a critical task for effective scheduling. A milestone that does not represent a meaningful product, accomplishment, or event to the performer is of very limited value.

23.5 Design

The goal of design is to develop a blueprint for successful production. From this perspective, design teams can be viewed as a factory whose product output is a production blueprint. The design factory must develop a roadmap to convert intellectual property, labor, and/or materials into a competitive product. This description of design is far broader than the conventional view of design as a process that produces an electrical circuit description or demonstration breadboard. Whether the product is a research report or a commercial wireless telephone, this design-factory concept can be used to some advantage.

As in the case of other manufacturing facilities, economy and efficiency can be obtained by loading the factory, but this loading will adversely affect the throughput time for a product and the ability to effect change. Since the goal of all productivity initiatives is to bring about cultural change, close examination of the loading of the design factory is required. If every engineer in the design factory is loaded with work accounting for a 110% workload factor, there will be little capacity to learn new concepts and methods. Yet, such learning is required to realize successful productivity initiatives. Change will come very slowly for a loaded design factory. If the design factory loading is reduced, efforts on new initiatives can fill the remaining capacity and change may occur more quickly. The more discretionary time that engineers have during the day, the more rapidly productivity changes can be made.

Important parameters to monitor for design factories include speed, cost, and value of the designs produced by the factory. Clearly, as in any other factory, access to required resources and raw materials is required to fuel the factory. For the design factory, these resources include appropriate office environments, office support, computers, design tools, literature access, and so on. An efficient design factory must also be maintained with appropriate support organizations. For many engineering applications, these organizations include marketing, quality assurance, reliability, and sales. An important concept unique to the design factory that should be considered when launching productivity initiatives is that the design factory improves its speed, cost, and value through learning. The design factory machinery (engineers) thrives on learning and teaching. This fact leads to some important concepts related to productivity initiatives.

We learn in many ways including through failure. Finding a failure quickly is often critical to the ultimate timely success. For the entire design factory to learn from failure, however, it requires that failures are identified and made public. There is a natural resistance to this in most engineering cultures, since failures have historically not been rewarded. It is important that the cost and value of failures be understood and that high value/low cost failures be rewarded. Early fast failures often stimulate the design machine to identify and apply appropriate resources to the most critical problems, thus developing successful designs more quickly. Rewarding such failure will create an environment that will favor fast design.

The need for learning also implies a need for intellectual stimulation and creativity. Diversity is a key ingredient to the development of a stimulating engineering culture. Diversity in both training and experience helps to fuel the creative machinery.

23.6 Productivity Metrics for Design-Earned Value

A design is only good to the extent that it helps accomplish the cost and schedule targets as well as the technical targets. Meeting performance and missing the cost or schedule can be a failure while relaxed performance at an earlier date and lower cost may provide success. Design metrics should provide a yardstick to measure cost, schedule, and performance variances during the design process. Earned value analysis (EVA) provides a simple method to develop design performance metrics, which indicate how the design factory is performing against its original plan.

To perform EVA, a baseline plan that includes the following is required:

- Planned budget to execute
- Planned schedule
- Detailed WBS
- Measurable milestones
- Actual budget to execute plan
- Actual schedule required to perform plan

Figure 23.2 illustrates an initial baseline plan for a design project prior to initiating work on any of the tasks. The example plan is composed of four separate tasks with estimated levels of effort of 80, 60, 100, and 50 h. The plan could be quantified in terms of monetary values or other measurable units instead of hours. The example plan is also divided into two distinct work periods. The x-axis on the chart is calendar time. The task hours may not correspond directly with the calendar time since task hours may represent part-time effort of an individual engineer or may represent several engineer's effort on a full-time basis.

The plan defines the Budgeted Cost of Work Scheduled (BCWS). BCWS is simply the budget or plan. The BCWS for an interval of time is the sum of the work of all tasks scheduled to be performed in that period. For example, in Figure 23.2, the BCWS for period 1 is the sum of 80 h for task A, plus 60 h for task B, plus 30 h of task C. Therefore, for period 1, BCWS = 170 h.

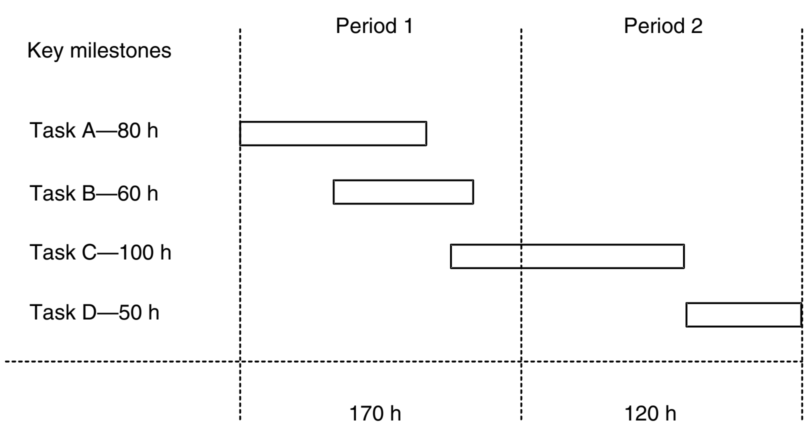


FIGURE 23.2 An illustration of an initial baseline plan for a design project prior to initiating work on any of the tasks.

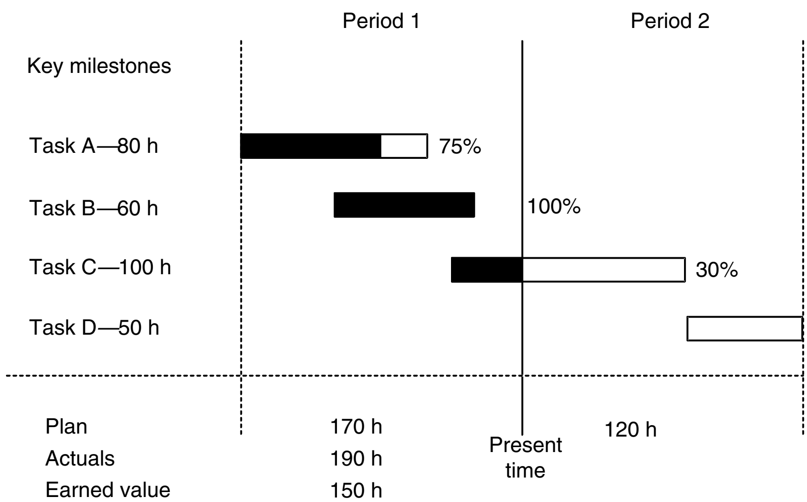


FIGURE 23.3 A representation of the progress completed on the initial baseline plan of Figure 23.2 through the end of period 1. In the figure, black bars are used to fill the work tasks to the extent that they have been completed to date.

The Budget at Completion (BAC) is also defined in terms of the plan and is simply the sum of all tasks for the entire project. For example, in Figure 23.2, BAC = 290 h.

As work progresses on the project, actual accomplishments and levels of effort can be compared to the baseline schedule. Figure 23.3 illustrates a possible representation of the progress through the end of period 1. In the figure, black bars are used to fill the work tasks to the extent that they have been completed to date.

The actual cost of work performed (ACWP) is the actual cost incurred for a task or period. From Figure 23.3, for period 1, ACWP = 190 h. This value is noted on the figure but is not represented in the graphics. Actual hours worked to achieve the progress reported is determined from other charging reports and simply recorded on the graph of Figure 23.3. The 190 h ACWP represents a variance from the baseline plan of 20 additional hours.

Figure 23.3 also indicates that the tasks originally planned for period 1 were not completed. Although tasks B and C were completed to the level indicated in the original plan, task A was scheduled for completion before the end of period 1 and is actually only 75% complete. The budgeted cost of work performed (BCWP) captures the earned value of the effort. The BCWP takes actual accomplishments and assigns the value that the original plan indicated it would cost to do this task. In Figure 23.3, BCWP for period 1 is obtained by summing up the costs identified by the black bars of each task. The earned value for period 1 is 150 h. This represents a deficiency from the baseline plan of 20 h.

From Figures 23.2 and 23.3 and the aforementioned definitions, several important performance metrics can be defined. Figure 23.4 illustrates some of these metrics graphically.

Schedule variance (SV) indicates whether the project is ahead or behind schedule. An SV can exist for several reasons and does not provide any indication about the cost performance. SV is defined as

$$\mathrm{SV} = \mathrm{BCWP} - \mathrm{BCWS} \tag{23.1}$$

The schedule performance index (SPI) indicates the schedule efficiency. SPI is the ratio of earned value to planned effort expressed as

$$\mathrm{SPI} = \frac{\mathrm{BCWP}}{\mathrm{BCWS}} \tag{23.2}$$

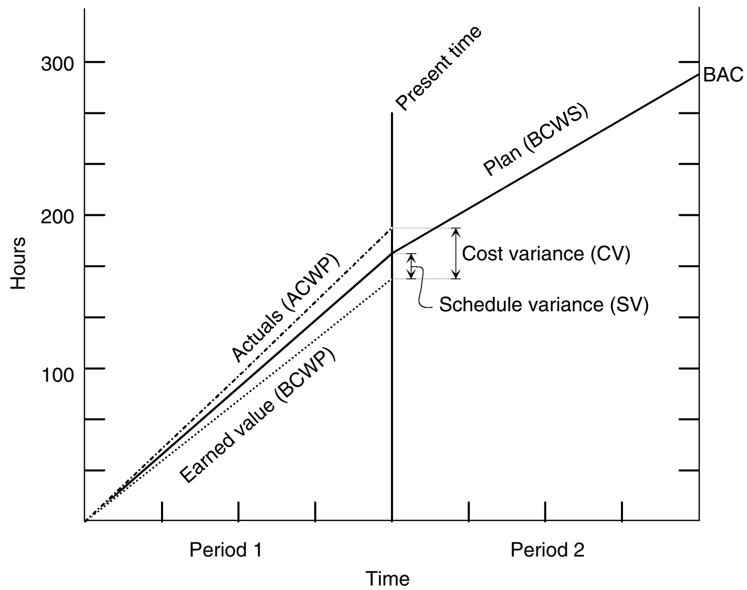


FIGURE 23.4 A graphical illustration of key metrics from an earned value program evaluation.

Cost variance (CV) indicates whether the project is over or below budget. A CV can exist for several different reasons and does not provide any indication about the schedule performance. CV is defined as

$$CV = BCWP - ACWP \tag{23.3}$$

Cost Efficiency is quantified with the Cost Performance Index (CPI). CPI is the ratio of earned value to actual effort expressed as

$$CPI = \frac{BCWP}{ACWP} \tag{23.4}$$

The Earned Value quantities defined earlier provide a method to evaluate design performance against the original plan. Evaluation of the CV and SV and efficiency is a first step to improving design performance.

A project that is on schedule and at budget will exhibit CPI and SPI values of unity. A low value for SPI indicates the project is behind schedule while a low value of CPI indicates that the project is over budget. Schedule and budget performance variance, however, can have many different root causes—including the use of a poor plan. It is important that the EVA is followed by an effort to determine the root cause of any poor performance metrics before any action to remedy the problem is adopted. For example, if schedule metrics are in deficit with the baseline plan and budget metrics are consistent with baseline plan, then this is an indication that the level of effort applied to the program is less than planned. Acquisition or identification of additional resources for the remainder of the program could lead to successful completion of the program (both budget and schedule goals met) for this case. If both schedule and budget metrics show a deficit from plan, there is an indication that the work is not on target for successful completion. Careful review of the original plan as well as the execution issues related to each task is required.

23.7 Manufacturing

The most widely known productivity initiatives focus on manufacturing metrics and yield analysis tools. These initiatives make extensive use of statistical process control. Statistical process control uses sampling and statistical methods to monitor the quality of an ongoing process such as a production operation.

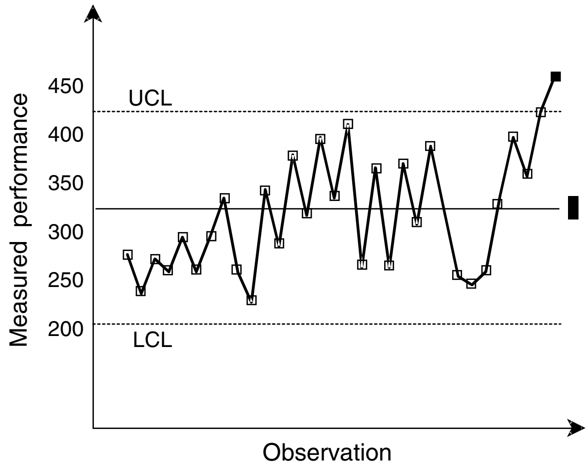


FIGURE 23.5 An example control chart.

A graphical display of statistical process variation known as a control chart provides the basis for determining when process variation is due to common causes (random variations) or to assignable, preventable causes. If our measurement equipment has enough accuracy, all processes will exhibit variations. The goal of reviewing control charts is to distinguish the variations intrinsic to the process from those that can be assigned a cause, and therefore can be controlled. An example control chart is presented in Figure 23.5. Control charts plot sequential evolution of time along the x-axis and the statistic that is being measured along the y-axis. Control limits (random variations inherent to the process) are statistically determined by careful observation of the process. These limits are plotted as bounds of expected process behavior. The mean value of the process is also plotted on the control chart. When variations exceed the control limits, the process is assumed to have shifted and adjustments in the process must be made. Variations within the control limits are assumed to be inherent to the process and therefore do not require process adjustments.

Control limits must represent actual process variability inherent to the process. Customer requirements or goals should never be used to establish control limits. To define control limits requires that an appropriate history of the process be collected.

23.8 Six Sigma

Six Sigma is the name applied to the Manufacturing Initiative that has become by far the most pervasive in the industry. This productivity initiative is a quality improvement and business strategy that was pioneered at Motorola in 1987. The Greek letter sigma designates the standard deviation from the average in a Gaussian distribution of any process or procedure. The common measurement unit of six sigma is "defects per million units." The sigma value of a process indicates how often defects are likely to occur. Three sigma means 66,807 defects per million. Four sigma translates into 6,210 defects. Six sigma quality implies only 3.4 defects per million. Ideally, product quality increases and cost decreases as sigma rises.

Although Six Sigma programs can be described in elaborate mathematical detail, the attainment of "six sigma quality" is not precisely defined. There are many ways to calculate defects per million units of a process. A unit, for example, can be almost anything: a system, a module, a component within a module, a part within a component, and so forth. Similarly, a process or procedure can be defined in many ways. The action of soldering a component to a board can be considered to be one single process, or can be broken down into several—picking the component, placing it, applying heat, and so forth. The choice of definition for a unit and a process can have a dramatic effect on the "defects per million" calculation.

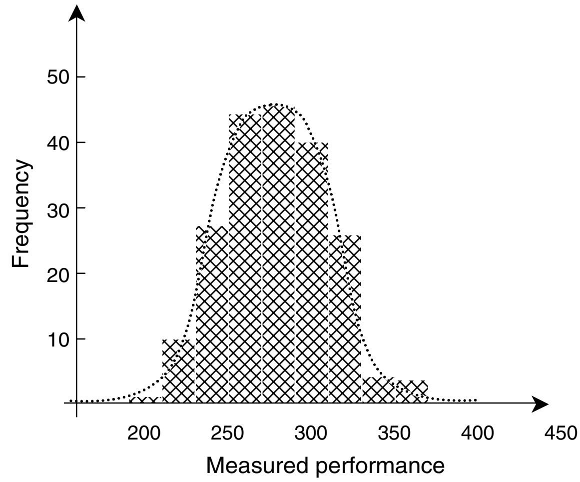


FIGURE 23.6 Example distribution statistics for a process. The observed statistics appear to approximate a normal distribution function.

When an auto manufacturer claims to have achieved six sigma production standards, it does not mean that they produce only 3.4 cars with some defect per million manufactured. The numeral “3.4” applies to the number of opportunities for errors. If a car consists of 10,000 different components and each one is assembled using 10 defined procedures, the opportunities for error totals 100,000 per auto. Six sigma production quality would imply 3.4 defects for every 10 cars.

Despite these ambiguities, six sigma programs have led to dramatic cost savings and competitive advantages to many companies that have embraced the philosophy. The list of companies that credit Six Sigma programs with significant cost and cycle time improvements is nearly endless.

Implementation of a Six Sigma program involves training employees to use statistical process control tools to achieve their defect reduction goals. Figures of merit are defined in terms of the statistical distribution of processed parts and these figures are monitored in an attempt to gain better control of the process.

Figure 23.6 illustrates the distribution statistics for a process. Plotted along the x-axis is the range of measured performance for the process of interest. The performance metric used for the process should be easily and quickly measured. This should be done in order to keep the measurement process from having a significant negative impact on the production cycle. Automation of the measurement and implementation of the measurement as a required part of the manufacturing process flow is important. The metric may not be a direct measurement of the actual desirable result, but it should reflect the quality and variability of the process (i.e., if the metric is controlled, the process should be controlled).

Plotted along the y-axis of Figure 23.6 is the frequency that a measurement was observed within the sample of process measurements. For a process that is dominated by random variations, the frequency of observation vs. measurement value plot is expected to describe a normal distribution. As seen from Figure 23.6, the normal distribution peaks at a specific (mean) value and drops off toward zero symmetrically in either direction away from the mean. The mean value of a distribution provides a good description of the central tendency of the process. A second important parameter describing a distribution is the standard deviation. Standard deviation provides a measure of the magnitude of the random variation inherent to a process.

23.9 Six Sigma Mathematics

The normal distribution function (also called the Gaussian distribution function after the German mathematician, Carl Friedrich Gauss) is the fundamental mathematical equation for Six Sigma metrics.

A normal distribution function is the indefinite integral of the normal density function, the graph of which exhibits a typical bell shape. The normal distribution function can be described by

$$F_{\mathrm{N}}(x') = k\int_{-\infty}^{x'} \exp\left[\frac{-(x-\mu)^2}{2\sigma^2}\right]\mathrm{d}x \tag{23.5}$$

where the constant k is called the normalizing coefficient and adjusts the amplitude of the function so that the integral of the distribution over all possible values of x will give a probability value of unity. The constant k can be expressed as

$$k = \frac{1}{\sigma\sqrt{2\pi}} \tag{23.6}$$

In Equations 23.5 and 23.6, μ is the mean, or average, of the distribution. For normal distributions, the mean value is also the point where the function reaches its maximum. The quantity σ is the standard deviation and characterizes the width of the distribution. For engineering processes, the mean value is typically the manufacturing target value and the sigma corresponds to the variability in the process. An ideal probability distribution would have a σ value that approached zero.

Two important metrics to analyze relative to six sigma performance are the C_{p} and C_{pk}. The C_{p} metric is referred to as the Process Capability Index and is given by

$$C_{\mathrm{p}} = \frac{\mathrm{CPU} + \mathrm{CPL}}{2} = \frac{\mathrm{USL} - \mathrm{LSL}}{6\sigma} \tag{23.7}$$

where CPU is defined as

$$\mathrm{CPU} = \frac{\mathrm{USL} - \mu}{3\sigma} = \frac{z_{\mathrm{U}}}{3}, \tag{23.8}$$

CPL is defined as

$$\mathrm{CPL} = \frac{\mu - \mathrm{LSL}}{3\sigma} = \frac{z_{\mathrm{L}}}{3}, \tag{23.9}$$

and the quantities LSL and USL represent the lower and upper specification limits. The quantities z_{U} and z_{L} that appear in Equations 23.8 and 23.9 are derived from normalizing the x-values of the observed measurement space to the process standard deviation using the transformation

$$z = \frac{x - \mu}{\sigma} \tag{23.10}$$

The C_{pk} metric is called the Process Bias and is similarly given by

$$C_{\mathrm{pk}} = \min\,(\mathrm{CPU}, \mathrm{CPL}) \tag{23.11}$$

For a process to be of six sigma quality, the following conditions must hold:

$$C_{\mathrm{p}} \geq 2 \tag{23.12}$$

and

$$C_{\mathrm{pk}} \geq 1.5 \tag{23.13}$$

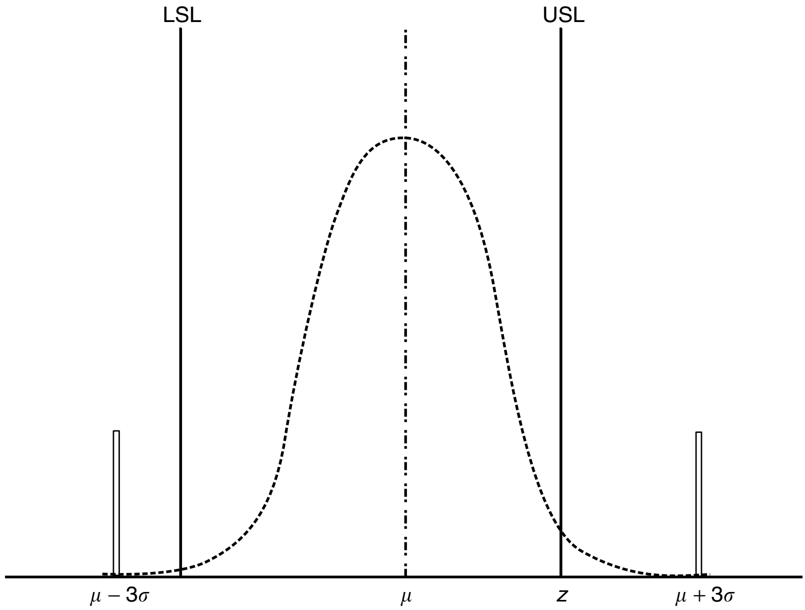


FIGURE 23.7 A process distribution chart with the quantities USL, LSL, μ, and σ illustrated on the plot. In this example, USL and LSL are not symmetric with respect to the mean while both are within the 6σ limits.

A process distribution chart is illustrated in Figure 23.7 with the quantities USL, LSL, μ, and σ illustrated on the plot. In this example, USL and LSL are not symmetric with respect to the mean while both are within the 6σ limits.

The Six Sigma goal may not be appropriate for every process as applied to every manufacturing task. The initiative to train all of the workers in a factory to monitor process variations and to understand the statistical metrics previously listed, however, almost certainly has merit. This training procedure and heightened awareness of yield and process variation issues may be the real value of Six Sigma Initiatives.

Bibliography

1. Hahn, G.J. and Shapiro, S.S., *Statistical Models in Engineering*. New York: John Wiley & Sons, 1967.
2. Montgomery, D.C., *Introduction to Statistical Quality Control*. New York: John Wiley & Sons, 1991.
3. Reinertsen, D., *Managing the Design Factory*. New York: The Free Press, 1997.
4. Womack, J.P. and Jones, D.T., "Beyond Toyota: How to Root Out Waste and Pursue Perfection," *Harvard Business Review*, September–October, 1996.

24

Cost Modeling

Leland M. Farrer
Farad Engineering

In the commercial industry, the ability to accurately estimate what a product can be built and sold for at a competitive profit, is a challenging task. Accurate cost modeling will simplify and organize the process into a powerful tool for the person entrusted with the task of quoting a project.

Each time a project manager is asked to quote on a future project, he is tasked with producing a cost for the development and cost for the end product. It does not seem to matter what industry you are in, the process is the same, and in the end, most managers succumb to estimating by experience rather than using a detailed costing model. The costing model is used to minimize the potential error by homing in on a model design that matches the market and, with feedback, can produce very good correlation with the end result.

The three major divisions of models include *product* focused, *service* focused, and *idea* focused. The classification of the three shows that the major emphasis is placed on *material*, *labor*, or *technology*:

1. Product-focused models are those where material costs are highest; where the added labor component has been minimized to the highest yield per hour spent. Examples include any production-line product that is automated.
2. Service-focused models are those where labor cost is highest. An example is a firmware product.
3. Idea- or technology-focused models involve an emerging technology with little track record. Examples are early life cycle inventions.

The cost model for all estimating has the following elements:

$$\text{Bill of materials} + \text{Process time} + \text{Yield margin} + \text{Overhead} + \text{Profit} = \text{COST}$$

The elements can be interconnected and developed into a spreadsheet application. Many organizations will have the company database connected to the cost model so that all elements are current and real time. A listing of the elements would be like this:

1. Bill of material (BOM) = Raw material cost (no labor for assembly or test) in $$
2. Process time = Operating cost per hour (labor) × Production cycle time = $$

3. Yield margin = (BOM + Process time)/Yield factor = $$
4. Overhead = (Process time + Yield margin) x Burden = $$. Overhead should also contain marketing costs as well as warranty or field returns as part of the burden element
5. Profit = (BOM + Process time + Yield margin + Overhead) × Percent of markup = $$
6. Cost = the $$ value is the total of all elements

Complimentary to the cost model is the risk assessment model. Where the cost model covers everything you know about, risk modeling evaluates and puts a dollar value on the things you don't know about that can surprise you. If risk modeling is done, it can be included with the other elements as a cost consideration.

24.1 BOM (Bill of Materials)

The bill of material is usually the total cost of all purchased parts such as resistors, semiconductors, etc., and fabricated cost such as PCB, shields, and outside fabricated parts. Depending on the internal requirement, all of these costs may be listed as raw materials with no handling markup. Some cost models will use this as the cost of goods for a product estimation and have a 20% handling charge on the dollar value as a burden element, independent of the G&A burden. This allows a softer reliance on hard yield and overhead elements later, but does mask the true cost. As with any production, the BOM costs will vary greatly with volume numbers. BOM costs at a quantity of 1000 will be much higher than at 1,000,000. Similarly, domestic BOM costs will be different than the same costs offshore.

24.2 Process Time

Depending on the item being estimated, the process time might be a resource output from a Gantt chart of all of the labor needed to design, test, and commission a prototype. This gets more difficult for development projects involving design review meetings and where many of the tasks are interrelated with critical path links. In complicated multiphase projects, the estimation of critical paths and margins to accommodate some slippage will become a requirement of the overall labor element of cost modeling. The cost element will be a total of all the individual worker hours at their unburdened individual rates.

Many models are available for setting up the production process. The decision regarding the method used will be based on industry needs and the maturity of the product.

In a manufacturing model, it might be the average cost per hour to operate a machine. The process time would then be the cost per hour times the cycle time to produce the item. Two types of process times must be considered each with a different cost impact. Variable (recurring) costs allow an incremental cost based on quantity changes. Fixed (nonrecurring) process cost is the base cost of holding the doors open (rent, etc.), and must be added. The final process time cost is the result of adding the percentage of use of the fixed cost and the increase due to volume increases.

24.3 Yield Margin

Yield margin is the amount of acceptable product versus the number that failed in the manufacturing process. In the semiconductor industry, it is the percentage or dollar value of good devices versus those units lost in wafer test, packaging, operational test, etc. In assembly, it is the number of units needed to set up the line, or parts lost in transport, QC, or customer rejection. If you need to deliver 100 units, and experience for like products have a total reject rate of 3%, then you will need a 3% margin (103) for proper yield. The cost, then, is the value of the rejected units spread across the 100 good units delivered.

24.4 Overhead

Overhead = Burden = $$(%). There is no easy number here. It must be calculated for each industry, geographical location, and will change as a company grows. For fledgling companies, this cost is shared by just a few individuals and can be quite high. If you are renting a location for \$2 to \$3 per square foot triple net, and have 5000 ft^2, this can be a significant number for 10 people. "Triple net" is a lease requiring the tenant to pay, in addition to a fixed rental, the expenses of the property leases such as taxes, insurance, maintenance, utilities, cleaning, etc. In this example, that is, \$10,000 to \$15,000 plus capital equipment amortization, insurance, utilities, facilities maintenance, etc., plus the administrative staff, secretaries, sales staffs, etc. These are items that are not directly adding labor to produce the product. If you are quoting on the development of a product and expect it to take 3 months of time and all of your direct labor people, then the overhead cost for the three months will be the number to use. The number can be proportionally less if only some of the resources are used. When commissions are used, it can be added as part of the burden factor, although it is more commonly added after the cost is computed and becomes part of the cost given to the customer. Usually when a company is stable in resources and size, the burden number can be expressed as a percentage of man-hours. If not, it must be computed on an individual basis of labor cost for each direct labor resource used.

24.5 Profit

Profit = (**BOM** + **Process time** + **Yield margin** + **Overhead**) × **Percent of markup** = **\$\$(%).** How much do you want to make? There are many ways to look at this number. If your market is highly variable, you may need a higher number to cover the low-income periods. If you are in the semiconductor industry, the number is part of a strategy for growth. Competition will also force some consideration of how much to charge. Generally the final number is a percentage of the cost of doing business and marketing opportunity based upon the customers' needs and what the market will tolerate. A point to remember for costing is the constant fluctuation of the exchange rates for foreign currency. Where the estimate covers a significant period of time, care must be taken to get professional assistance in this area. This could involve tying the value of the currency to the dollar or buying foreign funds to cover the exposure.

24.6 Cost

Now that we have put all of the elements together, the other side of the equation equals *cost*. It is now time to evaluate this number. Does the dollar amount meet the needs of the customer? Will the company be underbid by the competition? Are we leaving money on the table? If the answer to any of these is no, go back and see where the critical element is that will turn them into yeses. Remember, not all quotes are good ones. Some will not be good for you or your business. Most are good ones and it is up to you to find the right balance of value elements for your cost model equation.

Although this is not an exhaustive list of elements, it serves as a basis. Cost modeling is an extremely valuable tool in the quoting process, as well as a monitor of ongoing business health. Let us look at the major differences in the marketplace.

24.7 Product-Focused Models

When a technology has matured to the point where a high level of automation is implemented or where the percentage of material cost to labor contribution is high, the model can be very accurate and margins can be predicted to a high level of certainty. The BOM cost can be easily obtained from past purchases or the company's materials resource planning (MRP) database. The process time of the new design can be compared to the existing product and weighted accordingly. The yield and margins applied and the cost to produce can be generated from existing production runs. The chip manufacturing industry is an excellent example of this type of market, as are low-priced consumer market products. The material costs can be very accurately determined and process times are minimized as much as possible. In these markets, the volumes are high and per-unit costs are highly refined and generally low. This market is highly mature and cost models are finely tuned so that extracting the final profitability dollar amount has low risk. One key control for this model is feedback from the production line to confirm yields. Statistical process control (SPC) can be employed to measure performance and flag when corrective action is necessary. The run numbers from yield and process time allow confirmation of the model accuracy for future cost model estimating. With the high commitment of capital equipment and process time, cost modeling is necessary to minimize the loading risk. This model is well suited for fixed bid as well as projects with a time and material (T & M) cap.

24.8 Service-Focused Models

When the major contribution of the product is labor, the estimation is only as good as the reliability of the workers and their task execution. The elements of the model are the same, but the focus on past experience and more margin is required to compensate for this uncertainty. Examples of this are electronic prototype construction, software development, and warranty service. This model will display a wider swing in uncertainty relative to the product-focused model. Many service industries use this weighted model, and it is also taxed with individual performance estimating as a requirement. It is not helpful when worker substitution and different skill levels are factors, unless added margin is allowed.

24.9 Idea-/Technology-Focused Models

These can be the highest risk models. This is where invention is needed or relied upon for success. This can be seen in Gantt charts where the task "something magic happens here" is implied. It is also the area where the Intellectual Property (IP) (the idea) is not fully developed or needs refinement to make it ready for production. An example would be taking a reference design circuit and making an application-specific product. Due to the uncertainty of the elements of this model, the estimation is likely to be on the conservative side unless an investment in market penetration or a strategic alliance exists or is desired, in which case a more aggressive costing might be undertaken to secure the contract against competition.

Here again is the reliance on past experience. Regular reviews are needed of completed project performance to ensure that the model elements are operating properly. Failure to continually review and update the model elements on a daily, weekly, and monthly basis will result in an out-of-control erosion of the margins and profit.

24.10 Feedback

Probably the most important and often overlooked feature for accurate cost modeling is the use of feedback in the maintenance of the cost model. This can take several forms and is a featured process in all ISO 9000 documentation. Production feedback is used regularly on the factory floor at QA stations to ensure the quality of the product. No less important is the feedback from a development project to indicate the fiscal health of the team in meeting the milestones with calendar, material, and labor costs on track. This information not only needs to get back to the project managers for online project correction, but, also needs to get back to the contract administrators who generate the quotes to evaluate and adjust the cost model elements as necessary. Examples of corrections are the need for margin in the labor element if a unique skill set is required, or when vendor performance is in the critical path to meet a scheduled milestone and a calendar slip is likely. All feedback that will affect the performance accuracy must be routinely fed back into the cost model for correction.

24.11 Refinement of the Cost Model

In conclusion, the cost model is a living tool. Cost model effectiveness is only as good as the accuracy of the information going into each of the elements and the constant feedback to fine-tune it. A model that is good in a bull market may not be the one you need for the downturn era. As your company grows, the model will need to constantly evolve to keep pace with your needs. Putting the cost model into a spreadsheet will allow effective use and permit a lot of "what if" investigation as well as cost model evolution.

References

ECOM, European Space Agency (ESA) Cost Modeling, A general Description. http://www.estec.esa.nl/eawww/ecom/article/ecom.htm

Barringer and Associates, Life Cycle Cost Modelling, Freeware, http://www.barringer1.com/MC.htm

Skitmore, M., ed., *Cost Modelling.*

Cost Modelling (Foundations of Building Economics Series) by Martin Skitmore (Editor), Vernon Marston (Editor). Library Binding (2000).

25 Engineering Design Review Process

Leland M. Farrer
Farad Engineering

25.1 Overview

This chapter discusses the concepts involved in the implementation of a design review process for product development programs. The concepts also apply to efforts to refine an existing design review process in a total quality control (TQC) or total quality management (TQM) program. The reasons why a design review process is needed, key elements of such a process, and procedures for orchestrating an efficient program are outlined in this chapter.

25.2 Product Design Review as a Philosophy

In the process of doing business in the electronics industry, successful companies use a method of control that has been refined into a valuable tool. This tool is called the "Design Review." It requires a commitment to evaluation and feedback in the design process. Design reviews are the primary mechanism for control and success of the development process. In the TQC/TQM process, the details of how to do a design review are left to the company. Design reviews have many forms and applications to all industries. This tool has roots in statistical analysis, quality control, process control, financial analysis, and documentation control. When used properly, it helps make sure that resources of manpower and finance work hand-in-hand to the successful conclusion of product development. While, some would criticize that design reviews are time consuming and unnecessary, neglecting this process is likely to lead to surprises such as inappropriate technology usage, over-budget spending, technically impossible requirements ("magic happens"), or running out of calendar time.

It is necessary to validate and challenge the architectural design concepts, technology used, design margins, testing method, and labor and material cost estimates, as well as how to best meet the needs of the end user. This challenge brings out the best and the worst of the design in a controlled venue where corrective action can be taken on a proactive, constructive basis. It focuses the review with many perspectives, analyzes the design in a short time, and generates creative approaches that would otherwise not be considered. The theme is to bleed a little now rather than hemorrhage later. The rule is to avoid surprises, expose the risk, take action early, and be proactive. The development process is one of team commitment to success. The exhilaration of the group coming together to successfully solve an immediate problem is worth a little pain. This is an extremely dynamic process that can contribute to the individual professional growth of all participants. A very wise and compassionate mentor once said, "It is not necessary to find out who shot John. The fact that he was shot is enough. Now, let's make him healthy again." It is better to get a little bruised finding out that the team has a problem, early when it is fixable, rather than after it is too late to do anything constructive about it.

The design review process is not just used in the development part of product evolution. It plays an important role in the proposal stage all of the way through to design, new product introduction, production, yield analysis of production, field returns, and final retirement by product obsolesces. Each phase in the product life cycle needs monitoring and control to ensure success over time. The design review process fulfills that role in a formal controlled manner. In addition to these compelling reasons, the design review process also fits nicely in the requirements for ISO certification.

25.3 Product Development Documentation Sequence

Sometimes, in the rush to get a contract, a quote is made without involving the development project team. Besides putting a tremendous amount of stress on the team to do work to which they had no inputs, it also fails to provide an environment where good ideas and optimum technology choices can influence quotes in terms of time and money. The ideal contract proposal sequence is to precede proposal submission with a review. That review should explore ways to solve the problem and find possible gaps in customer specifications. Technological risks can be weighed and integrated into the proposal and quote. A design review at this stage allows the prospective team inputs into the proposal. From this process, we get the first elements of the product development documentation, acceptable customer specifications, and the architecture of the design solution. A timeline can then be determined using these elements. If the timeline does not meet the customer's schedule, alternative methods are best worked out during this initial phase. Team inputs and their "buy in" to the solution will help ensure success. On any project, the development of a detailed schedule is required. It defines the course from concept to conclusion. It also identifies project milestones and times when design reviews are warranted. A feasibility phase may be required by the customer in order to evaluate the project's probability of success. During this phase, technological suitability, patent issues, development costs, and availability of company resources are studied. Detailing the development process schedule and controlling decision making with design reviews will often alleviate customer concerns over proposed solutions.

25.4 Products Developed for the External Customer

In the final project proposal, the customer specifications and product description must be reasonable and practical to accomplish. Any customer requirements that require a "magic happens" step should be resolved as early as possible. Customer requirements are generally either *applications oriented* (uses existing hardware/software technology and has acceptable risk) or are *technology oriented* (uses new design creation or innovation and has higher risk). In either case, design reviews help ensure a full understanding of the product description and specifications.

Any contract generated to control project development must be clear and unambiguous. Decision points for progress approval should be built into the schedule and design reviews. Schedule progress is determined from engineers' design information and progress status. The design architecture for system's work should be firmed up in the proposal and quote stages. Component efforts require early decisions on circuit topology and manufacturing technologies. Device work also requires similar decisions about material and fabrication in the early phases of the contract effort. Engineering efforts must focus on details consistent with basic engineering plans. The goal is to make the design review report an argument for continued effort in the development process. Here, the customer may be an integral part of the decision process and can be a part of the design review team.

The scheduled design reviews reflect the orderly progression of effort. There is also a need to have nonscheduled design reviews for unexpected problems needing immediate attention. Anyone on the development team or related support group can request the project manager to have a nonscheduled meeting. The reason for the meeting may be anything from technology problems to the unavailability or seriously delayed delivery of a required component or outside process. The problem and a request for alternative solution presentations should be on the meeting agenda.

25.5 Internally Developed Products

For the development of internally used products, the lead entity is the marketing group with their fingers on the pulse of the market. Marketing's role is to identify customer's needs. Ideas may also come from a technological breakthrough in the engineering department. Involvement of both groups, marketing and development engineering, is just as important as with an outside customer for working up the internal product solution architecture. The product design goals and performance requirements fill the slot of customer specifications. Timelines and budget constraints are built into the schedule along with the milestones and design reviews, the same as with an external customer.

25.6 Design Review Preparation

The best preparation for a design review is to bring everyone up to a high level of understanding of the review agenda items. With complex design issues, one cannot expect someone to come to a meeting and contribute when they are not informed on the problem. Time to go over the design issue is needed before the meeting. Tables 25.1 through 25.6 list a typical range of required documents to support design review meetings for a communications device. It is by no means exhaustive.

Tables 25.1 through 25.6 provide examples of documents that might be needed for several different types of design reviews. Different departments supply the documents in the tables. Purchasing gets the

TABLE 25.1 Document Package for Proposal

Draft contract/proposal
Theory of operation
Time schedule
Budget
Risk analysis
FCC regulation compliance specifications
Experimental license application
Resource allocation
Customer specification
Customer product description

TABLE 25.2 Document Package for Proof of Concept (Breadboard Version)

Product description
Electrical schematics
Master costed parts list (bill of material)
Sole source components list
Test outline
Test specifications
Test results
Thermal calculations on dissipation
A working example of the product

TABLE 25.3 Document Package for Working Prototype (Typically Two PCB Iterations Are Needed)

Product description
Electrical schematics
Master costed parts list (bill of material)
Mechanical drawings
Sole source components list
Component layouts
Printed circuit artwork or files
Test outline
Test specifications
Test results including FCC compliance
Capital equipment list for testing
Test fixture
Working example of product

TABLE 25.4 Document Package for Pilot Preproduction

Product description
Electrical schematics
Draft costed parts list (bill of material)
Part numbers assigned to all parts
Mechanical drawings
Mechanical assembly drawings
Sole source components list
New material specifications
Component layouts
Silk-screen artwork
Printed circuit artwork or files
Test outline
Test specifications
Test procedures
Application for FCC certification if needed
Manufacturing performance margin analysis
Capital equipment list for testing
Test fixture
Working example of product

preliminary BOM (bill of materials) and costs it out. Drafting does initial drawings from customer product descriptions. Manufacturing engineers review test procedures for impact on the factory floor requirements for yield margins. Much of the company contributes to the information needs of the project. The design review documentation issue should be considered carefully so that *everything* relevant to the subject is collected before the meeting. Think of the meeting location as if it was a foreign county—bring everything!

TABLE 25.5 Document Package for Production Release

Product description
Product structure
Electrical schematics
Master costed parts list (bill of material)
Mechanical drawings
Mechanical assembly drawings
Special fabrication drawings
Sole source components list
New material specifications
Component layouts
Silk-screen artwork
Printed circuit artwork or files
Test outline
Test specifications
Test procedures
Capital equipment list for testing
Test fixture
Process sheet
Special instructions for QA
Working example of product

TABLE 25.6 For Software

Program logic flows to satisfy theory of operation
Preliminary simulation results
Timing charts
Regressive testing results
System testing strategy
Software program

There is nothing worse than to be in the trenches discussing solutions to the problem and to be missing an indispensable piece of information.

25.7 Project Design Review

The documentation requirements and deliverables may vary greatly in different industries and technologies, but design reviews, decision milestones, and deliverable requirements are always the main ingredients of a controlled development process. This process is controlled by the project manager. The project manager's role is to direct and ensure that the development process meets the schedule with adequate resources and effort. A regularly updated schedule keeps track of the status of the effort. Gantt charts are usually used to show task progress against real time labor and dollars spent. This graphical representation of tasks against the progression of time is the focus of the control process. It is a useful tool for planning, scheduling, and monitoring project progress. The Gantt chart provides a way to lay out the order and dependency of tasks to be completed. Figure 25.1 illustrates a basic Gantt chart for a hypothetical project.

An announcement preceding the design review meeting gives the time and date. It also informs participants of the reason for the review, who is invited, what documents are needed, and the agenda to be covered. Times and dates to circulate documentation prior to the review gives participants adequate time to read, digest, prepare questions, and request additional backup material.

A deliverable section in the contract details the elements of the final package. It tells the team when and how the project is complete. All interim design reviews are stepping stones to accomplish the deliverable goal. Documentation required in the final delivery must be in process throughout the development cycle.

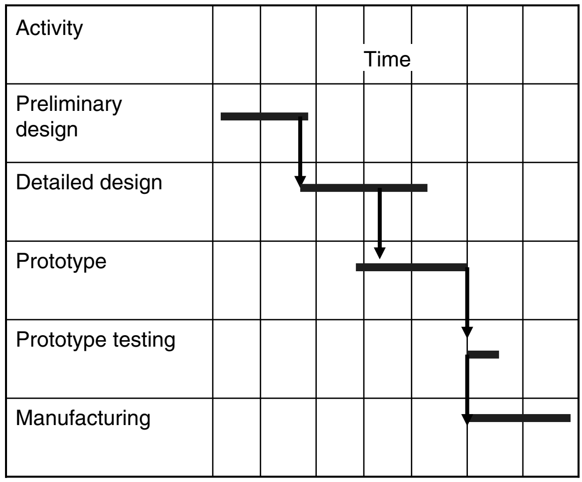


FIGURE 25.1 Example of a Gantt chart.

Items under revision control such as schematics, bills of material, testing procedures, software programs, and theory of operation are hammered out as development progresses. They are part of the elements referenced in each design review. With the architecture of the product defined in the proposal, the theory of operation presents an outline to work from. The progress of the engineers can be referenced to the desired theory for validation that they are on track to accomplishing the desired product solution. The final design review validates proof of performance results against specifications, compliance with the product description, theory of operation, and customer acceptance of the total physical product and documentation package. Figure 25.2 illustrates the typical flow of information for a design review.

25.8 Design Review Process

The design review can be for a milestone or the completion of a functional assembly. A meeting agenda for the design review is generated by the project manager. As mentioned previously, it identifies invitees, lists topic(s) to be covered along with the person responsible, and shows areas where supporting documentation is provided. An updated timeline of completed tasks in the schedule, resource status, and budget may also be provided. The status of the design development compared to the established schedule is provided by the lead development engineer. This includes software, hardware, and documentation. The status contains successful progress, problems encountered, conflicts, discovered risks, and so forth. The person responsible for each topic prepares the requested information and circulates it to the invitees listed in the agenda. All of this must be done well in advance of the actual meeting to allow participants to become familiar with the topic(s) prior to the meeting. Advance preparation makes the meeting a proactive results-oriented one instead of a teaching venue.

The actual meeting is a report-oriented presentation. While the project manger moderates, the responsible development engineer covers each item on the agenda. Discussions concentrate on the agenda topics, solutions to problems, cost and labor impacts to schedule, and so forth. A designated participant takes minutes in summary format. The minutes also include required actions and the people responsible for them. A single page should be sufficient for each meeting. The report summary restates the agenda items, lists the major and minor points of discussion in bullet format, identifies any action(s) necessary along with the person responsible, and gives a timeline for the completion of the action. Whether or not corrective action resolves a problem, the information is sent to the project manager who updates the action item list in preparation for the next design review. The project manager then takes whatever action is necessary to continue the development process. All results are distributed to the design review attendees.

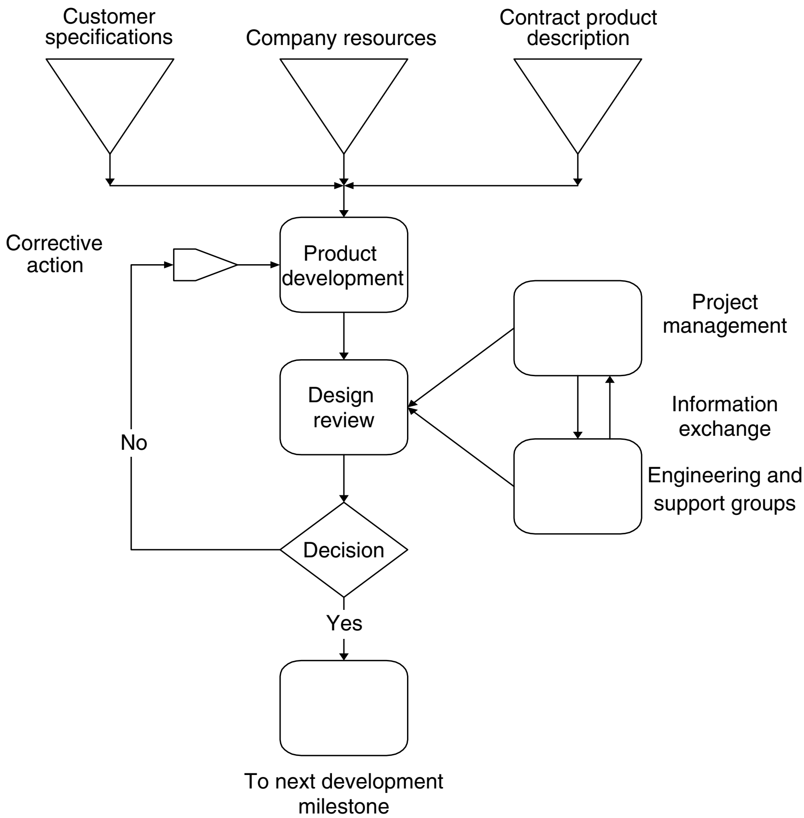


FIGURE 25.2 Design review process flow.

TABLE 25.7 Types of Design Reviews

Technology feasibility (proposal) review
Project (Contract) release review
Project progress (Milestone) reviews
Proof of performance review (Customer acceptance)
New product introduction review (Production release)
Production yield review (Cost reduction)
Warranty design review (Required design updates)
Design obsolesces review (End of life)
Other nonscheduled reviews (Problem is found)
Conflicts in specifications and negotiations strategy review

While this description may seem a bit involved and paper busy, the process focus is to prevent anything from being forgotten. The design review should not be mistaken for a catchall. It serves as a specific formal decision role. Regular informal development and/or status meetings are still needed and encouraged.

25.9 Types of Design Reviews

Table 25.7 shows some types of design reviews for a typical product development life cycle.

During the development phase of a product, conflicts in specifications often arise. A design review should detail the conflict, offer alternative solutions, determine implementation consequences, and trigger negotiations with the customer on any items with a compromise solution. The latter is best done after discussing thoroughly all of the facts, alternatives, and consequences.

25.10 Alternative Design Approaches, Fall Back Position Review

When the proposed action does not work out the way it was intended, a review to look into alternative approaches and recovery or fallback positions is called for. It can be the result of technology limitations where modifications to the specifications are needed. It can also be a result of a change in situation causing the cost to be out of budget limits. Patent infringement issues may surface and may need to be addressed. Any cost impact of a design alternative is a serious issue that requires full investigation of the source of the change to determine whether it is a one-time development cost or an increase to the proposed manufacturing cost.

25.11 Practical Design Review Outline and Report Form

The format of the design review outline varies with the industry and size of the company. There are also variations between hardware-oriented and software-oriented reviews. One form does not fit all of the applications. As long as the main elements previously discussed are incorporated into the format, each company and industry generates a form acceptable to their own needs.

25.12 Review of the Design Review Process

The review of the design review process falls under the TQC and product life cycle responsibility. Upon project completion, a review of the project from proposal to completion determines the effectiveness and efficiency of the team in meeting the customer's needs. This review compares the successes or failures of each development phase and whether the process needs to be altered to improve process procedure or execution. Changes in company size, economics, technology, competition, and customer base may also necessitate changes in the control process. Control processes are living documents. They must be reviewed and updated to remain effective and guarantee success in meeting both the company and customer needs.

25.13 Summary

The design review philosophy is one of control and feedback to enhance the development process and encourage personal growth in individual contributors. The size of the company has an impact on the process, but maintaining formal control should not be ignored. In many small companies, the principals wear multiple hats and tend to believe that formal design reviews are unnecessary. If the company desires growth, many expansion headaches can be eliminated by developing a sound control process with design reviews.

The design review process is a powerful tool for controlling the product development process. Only a few decades ago the process was workable but had far too much paper. Engineering firms spent too much time doing things manually. Even under these primitive conditions, the process was controlled and most projects were successful. Instant worldwide communications are now available making design reviews more efficient and effective. The design review process provides a means to challenge accepted thinking. Even though it can result in bruised egos, the process can provide professional growth for engineers and success for businesses.

26
Power Supply Management

Brent A. McDonald
Dell Inc.

George K. Schoneman
Rockwell Collins

Daniel E. Jenkins
Dell Inc.

26.1 Introduction

26.1.1 Short History (1970→2000)

1970—Switching frequencies were normally 50 kHz or less. Bipolar transistors were predominately used as the power switching elements [1].

1980—Switching frequencies were normally 100 kHz or less. Metal-oxide-semiconductor field-effect transistors (MOSFETs) were predominately used as the power switching elements. Control ICs were introduced. Constant frequency pulse width modulation became popular. Schottky diodes were introduced.

1990—Switching frequencies were normally 500 kHz or less. Current-mode control (CMC) became popular (peak CMC and average CMC). Synchronous rectifiers were introduced. Commercial-off-the-shelf (COTS) became popular. Zero voltage and zero current switching techniques were introduced.

2000–2006—Technology trends have not changed much since 1990. Typical switching frequencies are still 500 kHz or less. CMC remains on the forefront of many designs. Synchronous rectifiers are gaining popularity.

One area where there has been change is input power quality. Many countries around the world have adopted legislation mandating levels for power factor. This has led to the adoption of active and passive power factor correction (PFC) circuits in low cost, consumer electronics offline power supplies. The result has been a rapid growth in manufacture of PFC controller ICs. PFC has also enabled transition paths for new silicon technology such as silicon carbide (SiC) diodes.

The widespread use of the personal computer (PC), and its associated high current/low voltage requirements for the processor and memory, has led to the industry dominance of the multiphase DC–DC regulator. The high level of integration that became available allowed other control approaches, such as hysteretic, voltage mode, and dual slope to encroach on the dominance of the classic CMC.

Digital control is starting to make headway into the consumer power industry. Its ease of implementation, potential for lower parts count, and the ability to make changes in operation through software are enabling it to displace the traditional analog approach in select consumer electronics markets.

Future—Consumer PC voltages are projected to be as low as 0.5 V by the year 2010. These trends are suggesting dense hybridized power supplies colocated with the microprocessor (vertical integration).

The Automotive industry is going toward the more electric and all electric car.

26.1.2 Power Supply Types

Most power supplies are configured to supply a well regulated, constant voltage output, however, this does not have to be the case. Theoretically, a power supply can be configured to deliver constant current or power. Since most applications require some type of regulated voltage source, this topic will be addressed from that point of view.

Obviously, a power supply does not produce power for its load. It takes power from a source and reconditions the power so that the load can operate without damage or interference. If a power supply is required, it must be assumed that the power source is in some way not suitable to the device(s) it is powering.

This section will discuss three basic types of power supplies: linear regulators, charge pumps, and switch mode power supplies (SMPS). Each of these has an appropriate range of application and its own unique operational characteristics. A linear regulator takes a DC voltage with a higher magnitude than required by the load and steps it down to the intended level. The linear regulator typically accomplishes this by dropping the excess voltage across some kind of series pass element, commonly known as a series pass regulator. The second most common form of linear regulator is the shunt regulator. This approach often utilizes a fixed series impedance with a variable shunt impedance. Since the source voltage and load may vary, a feedback loop is usually used to minimize these affects on the output voltage. Figure 26.1 illustrates these two types of regulators.

A charge pump is capable of boosting a voltage to a higher level. A simplified charge pump is shown in Figure 26.2. These voltages may not be as well regulated as the SMPS or linear regulator approaches since their output regulation is dependent on items such as the forward voltage drop of the diodes and the initial tolerance of the power source being used. At turn-on, the inrush current may be very high. Essentially, the voltage sources are looking into a capacitor. The currents will be limited by the rate of rise of the source, the impedance of the diodes, and any parasitic resistance and inductance in series with the capacitor.

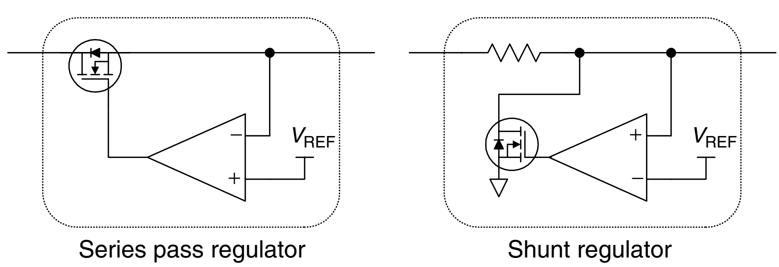


FIGURE 26.1 Two basic forms of linear regulators.

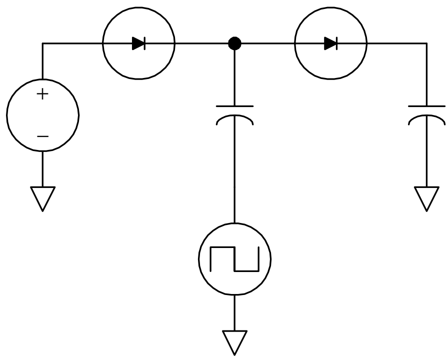


FIGURE 26.2 Typical charge pump.

Linear (Based on a 1 V low drop out device)	SMPS	Charge pump
No output voltage inversions possible	Output voltage inversions possible	Output voltage inversions possible
Low parts count	High parts count	Low parts count
Low design complexity	High design complexity	Low design complexity
Efficiency: 67–91%	Efficiency: 80–96%	Efficiency: 80%
Limited input voltage range	Wide input voltage range	Limited input voltage range

FIGURE 26.3 Power supply comparison (2 V $<$ V_{out} $<$ 10 V, 15 W).

It should be noted that by rearranging the circuit, the charge pump can operate as a voltage inverter. In other words it can generate a negative polarity output from a positive input, or a positive output from a negative input.

An SMPS operates by taking a DC voltage and converting it into a square wave and then passing that square wave through a low pass filter. In most applications, this allows for a more efficient transfer of power than that can be achieved with a linear.

For this discussion, it should be noted that the definition of DC voltage is not a well behaved, constant level input, but rather always the same polarity. Many SMPS circuits operate effectively with a pulsating DC input voltage. One example of this is the classic PFC stage. Its input is the full wave rectified output of the bridge rectifier.

Figure 26.3 lists a general comparison between these three types of power supplies.

26.1.3 Why Use a Power Supply?

Some power sources have voltages that vary 2:1 in normal operation and more than 10:1 during some transient conditions. Very few loads can function properly over these ranges and most will be damaged. A power supply can provide an interface to make these line voltage variations transparent, or at least tolerable, to the load's operation.

Some applications such as transmitters, stepper motors, processors, and so forth have large step loads. This may require lower output impedance than the source alone can provide. In other words, if the load requires a sudden burst of current, it is desirable that the output voltage remains in regulation. A low impedance capacitor bank can provide this step current. Alternatively, the control loop can sometimes be modified to supply such step loads.

26.2 System Issues: Specifications and Requirements

26.2.1 Input Voltage

This is an area where many mistakes are made in the initial design. The input voltage, or source, is rarely as well behaved as the title indicates. As an example, let us examine the average automobile electrical system as it exists in year 2006. The owner's manual will call it a "12 V" system, which sounds good, but when the alternator is charging, the voltage is typically $\sim$13.8 V ($\sim$15% increase). That is "typical," but actual numbers could range as high as 15 V ($\sim$25% increase). And it gets worse! Those voltages are defined as "steady state" or "normal operation." When heavy loads on the system are removed (headlights, starter motor, cooling/heating fans, air-conditioning, horn, etc.) the voltage on the bus can exceed 60 V (500% increase!) for durations $>$100 ms.

Standard household power is equally inhospitable. Lightning, power surges, and brownouts can result in a very wide input voltage range that your design may have to tolerate.

Battery operated units are somewhat better, but attaching and removing battery chargers can cause large voltage spikes. Owing to the internal impedance of the battery, transient step loads will cause degradation in the battery bus voltage.

26.2.2 Output

Power supplies are designed to power a specified load. In other words, there is no such thing as "one size fits all." For example, a supply that powers a 50 W load with 85% efficiency would be totally inappropriate for a 1 kW load.

A reactive load will introduce additional phase shift that a resistive load will not. This additional phase shift can create unstable or marginally stable systems. The module designer will often add significant capacitance to decouple one circuit from another. If this capacitance is large enough and not accounted for in the selection of the power supply, reduced stability margins, and/or a compromised transient performance can result. Likewise, if inductors are used to decouple the circuits, an inductive load can result and a similar detrimental effect on stability and transient performance is possible.

If the load exhibits large/fast step changes, the equivalent series resistance (ESR) of the output capacitors may cause the output voltage to drop below the valid operating range. The control loop will then try to correct for the error and overshoot can result.

A common misconception is that the DC accuracy of the output voltage is primarily a function of the reference. Usually, the output voltage is resistively divided down before being compared to the reference. If this is the case, the tolerance of the divider resistors needs to be accounted for. Offset voltages, offset currents, bias currents, and the location of the gains in the loop relative to the reference voltage can also significantly affect the DC settling point. This is all further complicated by the fact that these are all temperature-dependent terms.

The output ripple voltage waveforms of SMPS are harmonically rich. Generally, they will take the shape of a triangle or trapezoid. The fundamental frequency of these waveforms is at the switching frequency and will usually have harmonic content both above and below the fundamental. The content below the fundamental can be the result of input voltage variations, load variations, or injected noise. The ripple voltage specification listed on a data sheet or in a requirements document will probably give the maximum peak-to-peak voltage at the switching frequency or the root mean square (RMS) voltage. Often, larger spikes can be present that the ripple voltage specification does not account for. Figure 26.4 shows a typical output ripple voltage waveform and the frequency spectrum of that waveform. This particular case is a 20 V input, 3.3 V output, operating at 18 W.

26.2.3 Environment: Thermal, EMI/EMC, Efficiency

Looking at efficiency numbers can be misleading. Attention needs to be given to what the actual systems requirements are. In the case of a battery-operated system, efficiency may be of utmost importance. In this

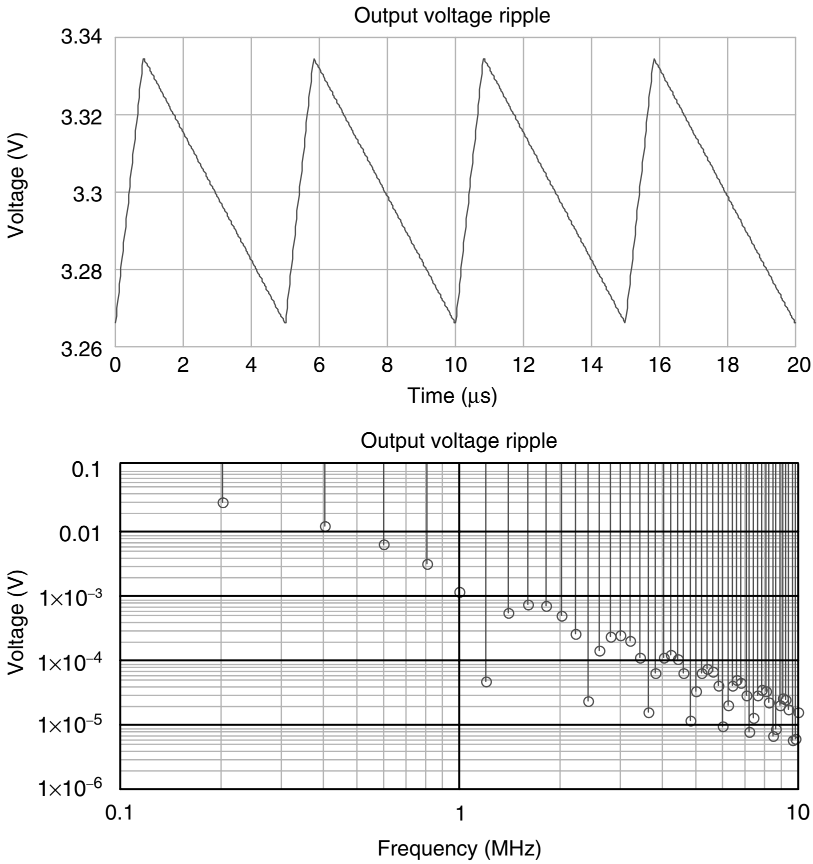


FIGURE 26.4 Waveform and spectrum of the output voltage of a 20 V-to-3.3 V power supply at a load of 18 W.

case, maximizing the time between battery replacements (or recharges) is a must. It may be particularly important to have some kind of ultra-low-power standby mode to prevent the needless waste of power when the supply is not being used. A system running off a generator or a low impedance power source probably is not as driven by the actual efficiency, but by the overall power dissipation. For example, it would not be uncommon to see a supply operate at 90% efficiency at the full-specified load. However, if that load drops by a factor of 10, the efficiency may drop into the 60% range. The reason for this is that the power supply usually requires some fixed level of overhead power. The larger the percentage of the actual load the overhead represents, the lower the efficiency. However, this is usually inconsequential. After all, if the source can provide the full load, it should have no problem powering a lighter load. Therefore, the efficiency at a light load may be irrelevant. The thermal loading of the box is what matters in this case.

But with the rising cost of energy generation and distribution, the drive for overall power supply efficiency improvement is gaining momentum. Several governmental agencies around the world are mandating minimum efficiency levels for AC–DC power supplies sold in their respective countries. In many cases, these regulations are specifying minimum efficiency across a wide load range. In one case, the specified load range is from 20% to 100% load. In these cases, the design must incorporate these requirements regardless of thermal constraints.

Electro magnetic interference/compatibility (EMI/EMC) can be divided into several categories: conducted emissions, radiated emissions, conducted susceptibility, and radiated susceptibility. A conducted emissions specification will define how much current can be put on the power lines for a given frequency. Figure 26.5 shows a typical input current waveform for a Buck derived SMPS. This particular case is a 20 V input, 3.3 V output, operating at 18 W.

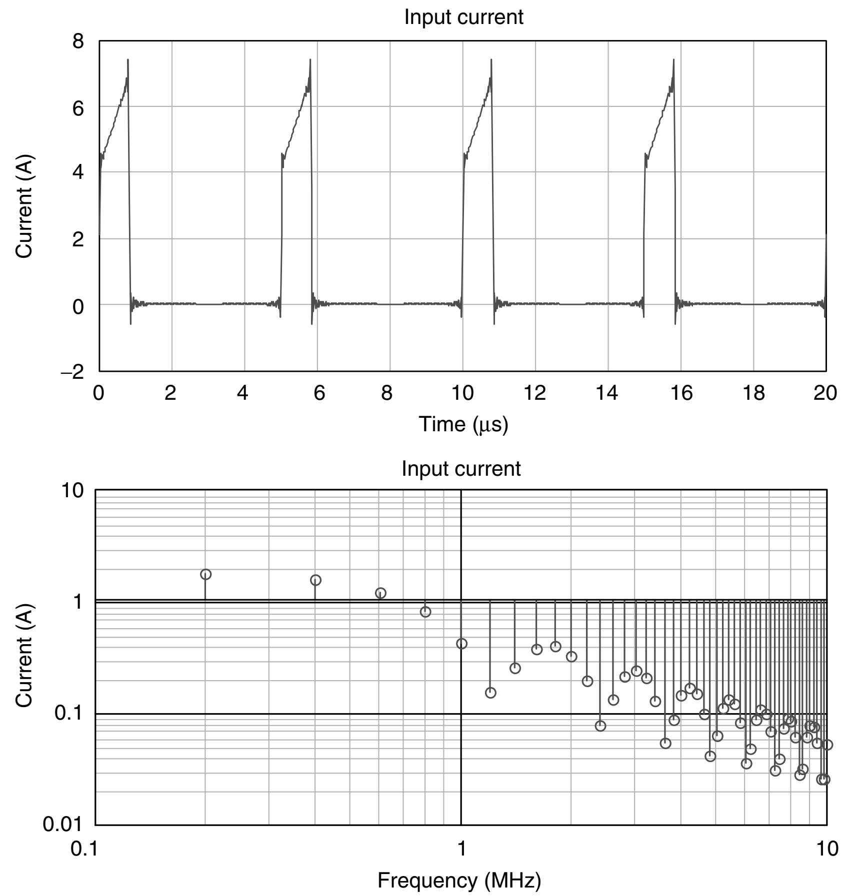


FIGURE 26.5 Waveform and spectrum of the input current of a 20 V-to-3.3 V power supply at a load of 18 W.

The EMI filter must attenuate this current spectrum per the system specifications. A two-pole LC filter is often used as a method for controlling emissions, similar to those shown in Figure 26.5.

A radiated emissions requirement will define the level of E and H field that can be radiated from the unit under test. A SMPS can be a significant source of these types of emissions, because of the high dv/dt and di/dt present. Typically speaking, if the conducted emissions are controlled, the radiated emission requirement is usually met. If necessary, shielded magnetics may be used to limit the H fields.

Conducted susceptibility defines the disturbance levels the power lines can be subjected to while the unit under test must continue to operate properly. Figure 26.6 shows an example of this kind of requirement.

In this case, the chart shows the RMS voltage that can be put on the power line for a range of frequencies. For example, the graph may start at 20 Hz, start to roll off at 2 kHz, and end at 50 kHz. (For a specific example see MIL-STD-461D (or 461 E) [2].) Many times, conducted susceptibility can be addressed by using an input-voltage-feed-forward term in the power supply control loop (this will be discussed in greater detail later).

Radiated susceptibility defines how much radiated E and H fields the unit under test can be exposed to and expected to operate properly.

26.2.4 Cost

Cost considerations are one of the most difficult areas to predict. So much of cost is dependent on the technology of the day. However, it can be safely stated that reducing size will increase both recurring

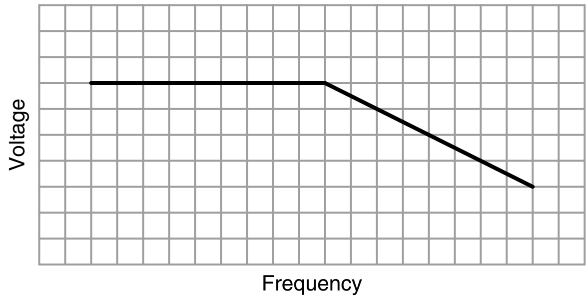


FIGURE 26.6 General conducted susceptibility curve.

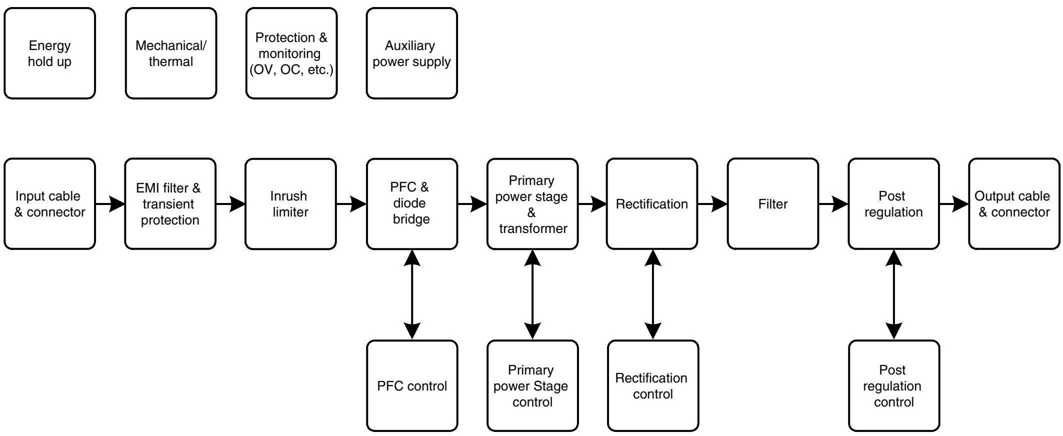


FIGURE 26.7 Power supply block diagram.

and nonrecurring cost. Increasing efficiency will increase nonrecurring cost and may or may not increase recurring cost. Typically speaking, the power inductor, bulk input and output capacitors, integrated circuits, and printed wiring boards will represent the majority of the cost.

26.2.5 Block Diagram

Figure 26.7 shows the basic blocks that make up a power supply. In reality, the blocks may be more integrated or arranged differently to accommodate various system requirements. For example, some systems have no requirement for energy storage or it is done in series with the power stage instead of in parallel.

The EMI filter and transient protection serve to protect the power supply from excessive line voltages and to ensure EMC. Many power lines have short duration high voltage spikes (e.g., lightning). Designing a power supply so that all of the relevant components can withstand these elevated voltages is usually unnecessary and will likely result in reduced efficiency and increased size. Normally, the energy contained in such a burst is low enough that a transient suppressor can be used to absorb it.

The inrush limiter serves to limit peak currents that occur due to the rapid charging of the input capacitance during start up or "hot swap." Hot swap is a special start up condition that occurs when a power supply is removed and inserted without the removal of input power. Once a power source is turned on, it can take several milliseconds before its voltage is up and good. This time reduces the peak current stresses due to the reduced dv/dt that the capacitors see. If a power supply is plugged into a source that is already powered up, significantly higher inrush currents can result.

Some power systems require the output voltages to be unaffected by short duration power outages (less than 200 ms for the purposes of this paper). An energy storage module accomplishes this. If the duration is longer, a UPS (uninterruptible power supply) may be used. In either case, some type of energy storage element (e.g., battery, capacitor bank) is charged and in the event of a loss of primary power, the supply runs off of the stored energy.

The power stage will be discussed in detail later.

The built in test (BIT) and monitoring circuitry typically monitors for elevated temperature, output over voltage, under voltage, and over current. Over voltage protection is necessary to limit the propagation of damage in a system. In order to do this, something needs to be done to clear the fault. Sometimes, a series switch is used to remove power from the supply. If this is the case, the switch normally latches open and can only be reset by cycling the primary power on and off. Another common method is a crow bar. A fuse or circuit breaker is placed in series with the power line. If an over voltage is detected, a short is placed on the output of the supply that will force the fuse to open. The down side of doing this is that it becomes a maintenance item, however, it does a good job of guaranteeing that the over voltage will be cleared.

Over current is also used to prevent the propagation of damage. Two common ways of addressing this are a brick wall current limit or foldback. Limiting the maximum current that the supply can put out makes a brick wall limit. For example, if the current limit is set at 6 A, the supply will never put out more than 6 A, even if a dead short is placed on the output. This works well, but sometimes the thermal loading during current limit can be more than the system can handle. In such a case, a foldback limit may be more appropriate. A typical characteristic is shown in Figure 26.8.

Notice that when the current hits the maximum level and the output voltage begins to drop, the current limit point also falls. This works to reduce the power dissipation in the supply. Care needs to be used when working with foldback limiting. It is possible to get the supply latched in a low voltage state when there is no fault.

Under voltage is used to prevent erroneous operation. Often a signal is sent out to report an under voltage condition. This signal can be used to hold a processor in reset or to inhibit another operation. Latching a supply off on an under voltage can be problematic, due to the danger that the supply may never start up!

In an effort to reduce size, many SMPS converters rely on a transformer as an easy way to produce multiple outputs. The basic idea is for one of the outputs to be tightly regulated through a control loop and the auxiliary outputs to track by transformer turns ratio. The problem is variations in the inductor direct

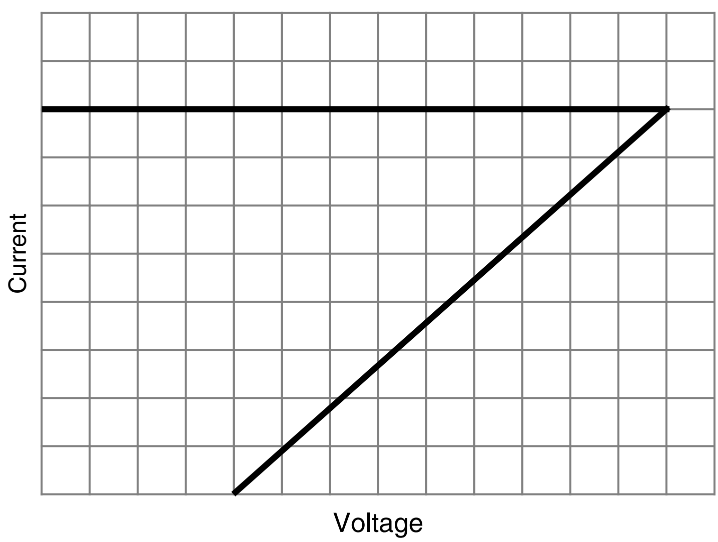


FIGURE 26.8 General foldback current limit curve.

current resistance (DCR), diode forward voltage drop, transformer coupling, load, temperature, and so forth will all cause the auxiliary voltages to vary. This additional variation appears on top of the initial tolerance established by the control loop. The bottom line is the auxiliary windings may not have sufficient accuracy. If this is the case, there are post regulation methods that can tighten the output tolerance with a minimal efficiency hit.

26.3 Power Supplies

Soft start is an area that can often be a problem. Figure 26.9 illustrates this problem.

When a power supply is commanded to turn on, the error amplifier will attempt to raise the output voltage as quickly as possible. Once the output voltage is in regulation, the error amplifier must slew quickly to avoid overshoot. The problem is compounded by the fact that stability considerations demand that the error amplifier be slowed down. This usually results in an over voltage.

Active circuitry can be added to force the error amplifier's output to rise slowly to the level required for proper regulation. There are a variety of ways to accomplish this, and it is essential that it be addressed in most systems.

When talking about SMPS, it is customary to talk about duty cycle (herein referred to as D). This is the percentage of time that the main switch is on per cycle. The remainder of the cycle is often referred to as 1-D (or in the literature D′) (Figure 26.10).

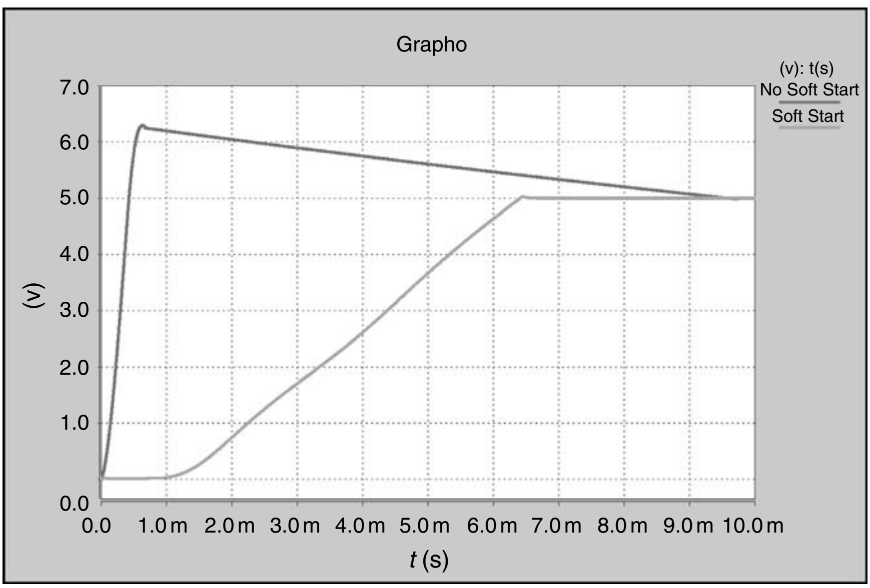


FIGURE 26.9 Typical start up characteristics with out Soft Start.

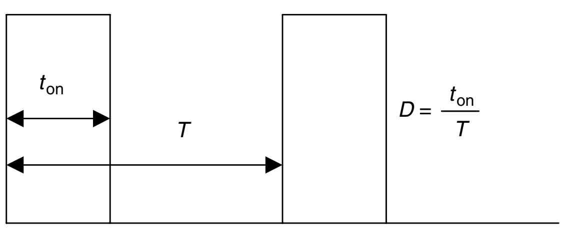


FIGURE 26.10 Duty cycle definition.

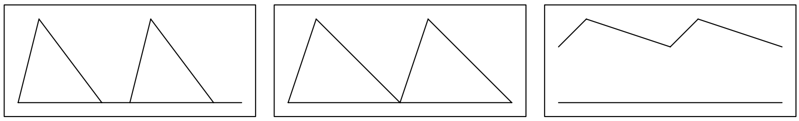


FIGURE 26.11 Current waveforms of discontinuous conduction mode (DCM), critical conduction, and continuous conduction (CCM).

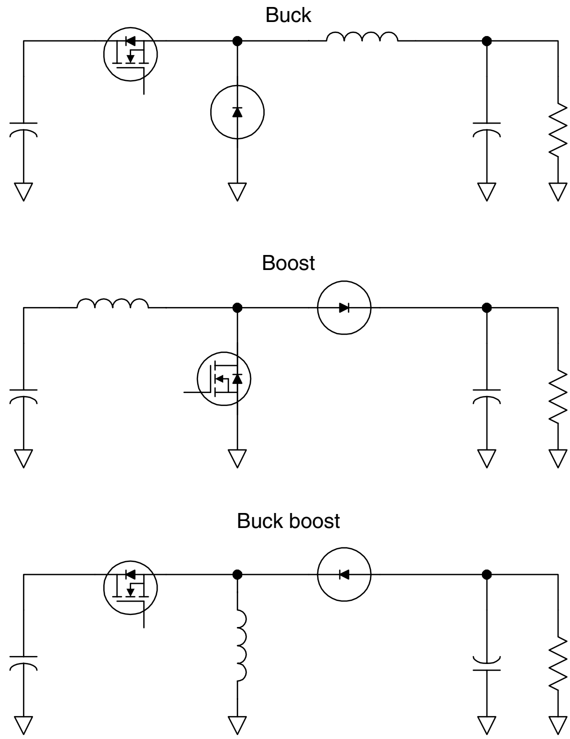


FIGURE 26.12 Basic nonisolated DC–DC converter power stage topologies.

SMPS are often categorized according to the state of the inductor current. Figure 26.11 shows the inductor current for a simple Buck converter in steady-state operation. The first picture shows discontinuous conduction mode (DCM). This condition occurs when the inductor current goes to zero during a portion of the 1-D time. A special case exists when the inductor current just reaches zero. This case is shown in the second picture and is called critical conduction. The third picture shows continuous conduction mode (CCM). This condition occurs when the inductor current is greater than zero throughout the cycle.

Figure 26.12 shows the power stages for the basic nonisolated DC-to-DC switching power supply topologies. In each case, there are two switches (MOSFET, diode), an inductor and an output capacitor. The input capacitor is shown to represent the low impedance input required for operation. The resistor on the output models the load, however, the load may not necessarily be resistive.

The CCM Buck converter is the simplest to understand of the three approaches. It takes an input voltage and steps it down. Essentially, the MOSFET and diode create a square wave at the input to an LC filter. The high-frequency content of the square wave is filtered off by the LC, resulting in a DC voltage across the output capacitor. Since the AC portion of the inductor current is the same as the output capacitor current, it is relatively easy to reduce the output ripple voltage by increasing the inductance. Figures 26.13 and 26.14 show the MOSFET and diode currents, respectively. Please note that these waveforms have an identical shape for each of the three topologies.

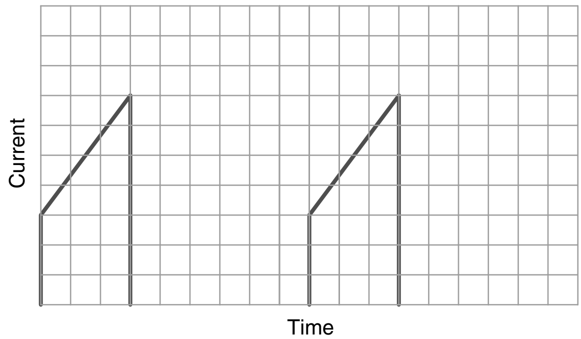


FIGURE 26.13 MOSFET current.

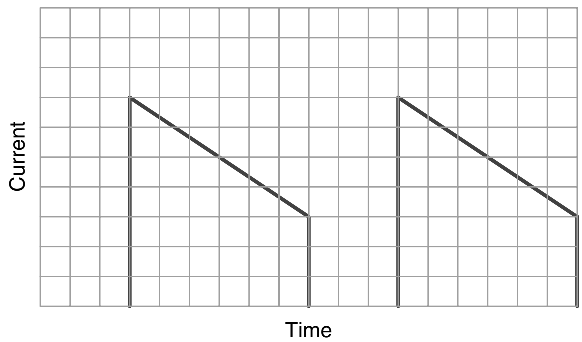


FIGURE 26.14 Diode current.

Duty cycle	Buck	Boost	Buck-boost
CCM	$\frac{V_{out}}{V_{in}}$	$\frac{V_{out}-V_{in}}{V_{out}}$	$\frac{V_{out}}{V_{out}-V_{in}}$
DCM	$V_{out}\cdot\sqrt{\frac{2\cdot L\cdot f_s}{R\cdot V_{in}\cdot(V_{in}-V_{out})}}$	$\frac{1}{V_{in}}\cdot\sqrt{\frac{2\cdot L\cdot f_s\cdot V_{out}\cdot(V_{out}-V_{in})}{R}}$	$\frac{V_{out}}{V_{in}}\cdot\sqrt{\frac{2\cdot L\cdot f_s}{R}}$

FIGURE 26.15 Duty cycle definitions.

The CCM boost converter operation is a little more complex. This time the MOSFET shorts the inductor to ground, allowing the inductor current to ramp up (Figure 26.13 is the MOSFET current). When the MOSFET turns off, the inductor current continues to flow by turning the diode on (Figure 26.14 is the diode current). For equal positive and negative volt-seconds (the product of voltage and time) to be maintained across the inductor, which is necessary to prevent saturation of the inductor, the output voltage must be higher than the input voltage. The equal volt-second constraint must be satisfied in any topology. The CCM buck-boost (or flyback) operates in an energy transfer mode. When the main MOSFET closes, the inductor charges. This charge action is shown in Figure 26.13. Once the MOSFET opens, the current in the inductor continues to flow and turns the diode on, transferring charge to the output capacitor. In the noncoupled approach, this always results in a voltage inversion on the output. Again, this stems from the equal volt-second constraint. Figure 26.14 shows the diode current.

The duty cycle required to generate a given output voltage is different in each topology and is dependent on the mode of operation. When in CCM, D is defined only in terms of input and output voltage. If in DCM, the relationship is more complex. The main point here is in CCM, the output voltage is independent of load. In DCM, it has the additional dependence on inductance, switch frequency, and load (Figure 26.15).

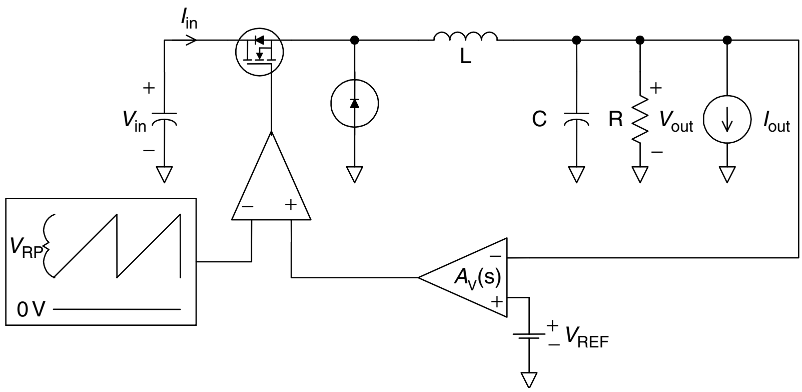


FIGURE 26.16 Basic nonisolated Buck with voltage mode control (VMC).

$$\begin{bmatrix} \dfrac{V_{in}}{I_{in}} & \dfrac{I_{in}}{I_{out}} \\[2ex] \dfrac{V_{out}}{V_{in}} & \dfrac{V_{out}}{I_{out}} \end{bmatrix} = \begin{bmatrix} \dfrac{s^2 + \dfrac{s}{R \cdot C} + \dfrac{1}{L \cdot C}\left(1 + \dfrac{v_{in}}{V_{RP}} \cdot A_v(s)\right)}{\dfrac{d^2}{L} \cdot \left[s + \dfrac{1}{R \cdot C}\left(1 - \dfrac{v_{in}}{V_{RP}} \cdot A_v(s)\right)\right]} & \dfrac{\dfrac{d}{L \cdot C \cdot R} \cdot \left[s \cdot L \cdot \dfrac{v_{in}}{V_{RP}} \cdot A_V(s) + R \cdot \left(1 + \dfrac{v_{in}}{V_{RP}} \cdot A_v(s)\right)\right]}{s^2 + \dfrac{s}{R \cdot C} + \dfrac{1}{L \cdot C} \cdot \left(1 + \dfrac{v_{in}}{V_{RP}} \cdot A_v(s)\right)} \\[4ex] \dfrac{\dfrac{d}{L \cdot C}}{s^2 + \dfrac{s}{R \cdot C} + \dfrac{1}{L \cdot C} \cdot \left(1 + \dfrac{v_{in}}{V_{RP}} \cdot A_v(s)\right)} & \dfrac{\dfrac{-s}{C}}{s^2 + \dfrac{s}{R \cdot C} + \dfrac{1}{L \cdot C} \cdot \left(1 + \dfrac{v_{in}}{V_{RP}} \cdot A_v(s)\right)} \end{bmatrix}$$

FIGURE 26.17 Voltage mode control (VMC) transfer functions.

Figure 26.16 shows one of the simplest standard methods for regulating the output voltage of a SMPS, voltage mode control (VCM). For simplicity, the Buck typology in CCM has been chosen.

A constant duty cycle applied to the MOSFET switch will produce a DC output voltage equal to the input voltage multiplied by duty cycle. If the duty cycle goes up, the output voltage will go up. If it goes down, the output voltage will also go down. The device driving the MOSFET is a comparator. The inputs to the comparator are a signal proportional to the output voltage (V_e) and a voltage ramp. If V_e goes up, D goes up and raises the output. Likewise, if V_e goes down, D goes down and the output falls. Some type of compensation is placed between the output and V_e to close the control loop [labeled $A_v(s)$ in the diagram]. This compensation will control the stability and transient performance of the supply. Figure 26.17 contains the mathematical definition of several transfer functions for the above system.

These equations are based on a modeling technique called state space averaging [3]. Practically, they only apply to 1/10–1/3 of the switch frequency. They are based on an averaged linearized model of the power supply. They do not contain valid information about the large signal response of the system, only the small signal. When this method is applied to a DCM topology, the standard approach may need to be modified to obtain a more accurate result [4]. If the voltage ramp that is fed into the comparator is proportional to the input voltage (generally referred to as input voltage feed forward), these equations simplify to the following (Figure 26.18).

The most practical result of this simplification is that perfect input transient rejection is achieved. In other words, the output will not be perturbed by variations on the input. Please keep in mind that these equations are simplified and idealized. Perfect input voltage transient rejection is not possible. Some input variations will reach the output due to delay times, parasitic effects, and so forth.

$$\begin{bmatrix} \dfrac{V_{\text{in}}}{I_{\text{in}}} & \dfrac{I_{\text{in}}}{I_{\text{out}}} \\[2ex] \dfrac{V_{\text{out}}}{V_{\text{in}}} & \dfrac{V_{\text{out}}}{I_{\text{out}}} \end{bmatrix} = \begin{bmatrix} -\dfrac{R}{d^2} & \dfrac{\dfrac{d}{L\cdot C\cdot R}\cdot\left[s\cdot L\cdot A_{\text{v}}(s)\cdot K + R\cdot\left(1 + A_{\text{v}}(s)\cdot K\right)\right]}{s^2 + \dfrac{s}{R\cdot C} + \dfrac{1}{L\cdot C\cdot R}\cdot\left(1 + A_{\text{v}}(s)\cdot K\right)} \\[4ex] 0 & \dfrac{\dfrac{-s}{C}}{s^2 + \dfrac{s}{R\cdot C} + \dfrac{1}{L\cdot C\cdot R}\cdot\left(1 + A_{\text{v}}(s)\cdot K\right)} \end{bmatrix}$$

FIGURE 26.18 Voltage mode control (VMC) transfer functions with input voltage feed forward.

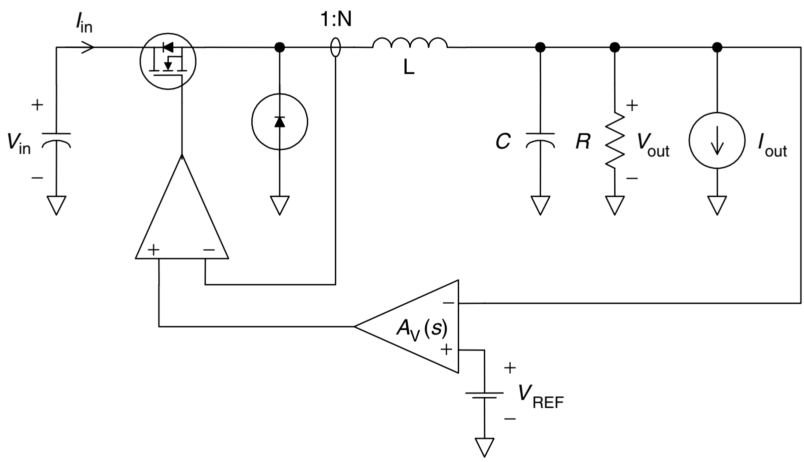


FIGURE 26.19 Basic nonisolated Buck with peak current-mode control.

Figure 26.19 shows one of the most popular ways to control a power supply, peak CMC. The fundamental difference between CMC and VMC is that the ramp used for comparison to V_e is now directly proportional to the inductor current. This action creates an inherent feed-forward term that serves to reject input voltage transients. However, it does not provide the ideal rejection that the VMC with input voltage feed-forward does. This can be achieved by adding slope compensation. If a constant slope (exactly equal to ½V_{out}/L) is added to the ramp, derived from the inductor current, the ideal input transient rejection is achieved [5]. Again, this is an ideal approximation. In reality, delay times and circuit parasitics will allow some of the input voltage variations to reach the output.

The following are the linearized state space averaged equations with the slope compensation equal to ½ the down slope of the inductor current (Figure 26.20). The technique used for modeling the current loop does not model the sub-harmonic oscillation described widely in the literature. Models that are more complex are needed to predict this effect [6]. (Please note that R_s/N is the gain of the inductor current to the input of the comparator and f_s is the switching frequency of the converter.)

26.3.1 Synchronous Rectifiers

Synchronous rectifiers are generally used to improve power conversion efficiency, particularly for low-output voltage converters. A low ON resistance electronic switch, typically an n-channel MOSFET, is placed in parallel with (or in place of) the rectifier diode(s). The forward diode drop of a normal rectifier diode is in series with the output voltage, so as lower and lower output voltages are required by modern processors (some core voltages are now under 1 V) and other ICs (DSP ICs routinely require voltages of 1.8 V or less) the losses in the rectifiers become a dominant factor in overall efficiency. Even a Schottky

$$\begin{bmatrix} \dfrac{V_{\text{in}}}{I_{\text{in}}} & \dfrac{I_{\text{in}}}{I_{\text{out}}} \\[2ex] \dfrac{V_{\text{out}}}{V_{\text{in}}} & \dfrac{V_{\text{out}}}{I_{\text{out}}} \end{bmatrix} = \begin{bmatrix} -\dfrac{R}{d^2} & \dfrac{\dfrac{d}{R\cdot C}\cdot\left(\dfrac{2\cdot L\cdot A_v(s)\cdot f_s\cdot N}{R_s}+d\right)\cdot s+\dfrac{d}{L\cdot C}\cdot\left[1+d+2\cdot f_s\cdot L\cdot\left(\dfrac{A_v(s)\cdot N}{R_s}-\dfrac{1}{R}\right)\right]}{s^2+\left(\dfrac{1}{R\cdot C}+2\cdot f_s\right)\cdot s+\dfrac{1}{L\cdot C}\cdot\left[1+d+2\cdot f_s\cdot L\cdot\left(\dfrac{A_v(s)\cdot N}{R_s}+\dfrac{1}{R}\right)\right]} \\[4ex] 0 & \dfrac{\dfrac{-1}{C}\cdot\left(s+2\cdot f_s\right)}{s^2+\left(\dfrac{1}{R\cdot C}+2\cdot f_s\right)\cdot s+\dfrac{1}{L\cdot C}\cdot\left[1+d+2\cdot f_s\cdot L\cdot\left(\dfrac{A_v(s)\cdot N}{R_s}+\dfrac{1}{R}\right)\right]} \end{bmatrix}$$

FIGURE 26.20 Transfer functions of a basic nonisolated Buck with peak current-mode control with optimal slope compensation.

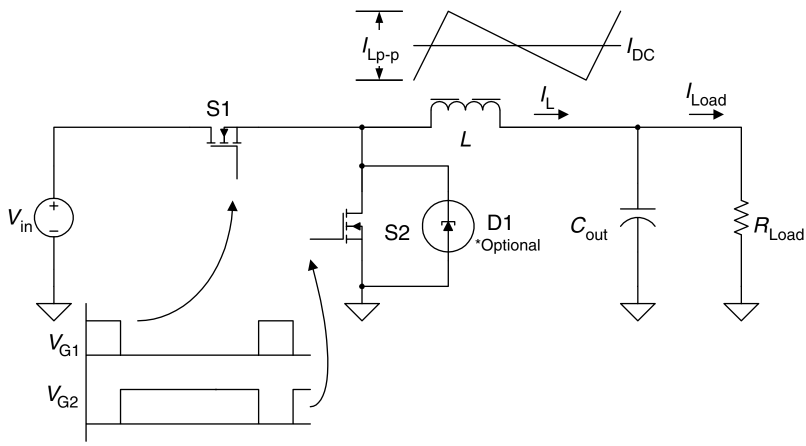


FIGURE 26.21 Buck converter with synchronous rectifier.

rectifier with a voltage drop of under 0.4 V is a major loss factor for a 3.3 V DC or lower output voltage. For example, for a 12 V DC input Buck converter generating a 1 V DC output voltage the efficiency is limited to 73% , even considering no losses other than diode conduction loss. This is partially due to the fact that the rectifier diode (commonly called the "catch" diode) is conducting over 90% of the time when the step-down ratio is large (12 V to 1 V). The efficiency is affected less with lower input voltages, since the step-down ratio is not as high, but rectifier conduction loss is still the dominant loss factor. Figure 26.21 shows a basic nonisolated Buck topology DC–DC converter with a synchronous rectifier placed in parallel with the normal Schottky rectifier diode.

In a MOSFET power switch the voltage drop across the drain to source terminals is proportional to the current flowing in the part. The voltage drop across the part is equal to the current times the ON resistance of the part ($R_{\text{ds(ON)}}$). Improvements in MOSFET fabrication technology have resulted in lower ON resistance and lower gate charge (gate charge determines how much power is lost in turning the MOSFET on and off at a given frequency and drive voltage) parts. Low voltage MOSFETs ($V_{\text{d-s(max)}} \leq 30$ V) are now available with ON resistances below 5 mΩ. For a 10 A current, the corresponding voltage drop is only 50 mV, as opposed to 0.4 V in the Schottky diode example given above. For the 1 V DC output voltage example given above, synchronous rectification can improve efficiency from 73% (with a Schottky rectifier) to the mid to upper 80% range, dependent on the load current and MOSFETs used. This significantly decreases the amount of heat that must be managed. An efficiency improvement from 73% to 85% reduces power loss in half, thereby reducing heat sink requirements, resulting in size and weight savings. Cost reduction due to decreased cooling needs must be traded-off against any cost increase due to increased complexity of changing a rectifier diode to a synchronous rectifier MOSFET power switch.

The need to provide synchronous rectifier MOSFET gate drive is a significant part of the trade-off. Power MOSFETs require gate drive power proportional to the total gate charge (Q_T—usually in nanoCoulombs—nC) of the device, the gate drive voltage needed to properly enhance the MOSFET and the switching frequency. The basic formula is gate drive power $= 2 * Q_T * V_{g-s} * f_s$. The numeral "2" in the equation accounts for the fact that, in a normal hard-switched gate drive circuit, equal power is dissipated during turn-on and turn-off of the MOSFET. Most of this power is dissipated in the ON resistances of the gate driver circuit. The power needed to turn the synchronous rectifiers on and off must be considered when calculating the efficiency difference compared to using a rectifier diode. As a general rule, lower $R_{ds(ON)}$ parts have higher gate charge numbers. The important trade-offs are whether the additional gate drive power and added circuit complexity are sufficiently offset by the reduction in conduction losses from the lower forward voltage drop in the synchronous switch. Load current, load current variation, voltage conversion ratio (input voltage to output voltage), and switching frequency all affect this trade.

A synchronous MOSFET switch does not necessarily replace the Schottky rectifier diode. To prevent potentially destructive (and very lossy) cross conduction of the high side MOSFET and synchronous rectifier MOSFET some dead time must be allowed between turn-off of one switch and turn-on of the other. However, current must continue to flow in the filter inductor so the parasitic diode in the synchronous MOSFET will act as a catch diode. This diode has a large forward voltage drop and is slow to turn-off. For maximum efficiency (typically a 1–2% improvement), a Schottky diode is placed across the synchronous MOSFET. This diode can usually be smaller than if it were conducting full time since it only conducts during the overlap time before the synchronous MOSFET gets fully enhanced (fully turned on) and, therefore, its average power dissipation is low. For it to be effective it must be placed physically as close as possible to the MOSFET, otherwise parasitic inductance will hinder its operation. This becomes more critical at high switching frequencies, since the dead time becomes a higher percentage of the on time of the switch.

MOSFET synchronous rectifiers can be driven, either by an active logic controlled drive circuit (i.e., usually incorporated into the control IC), or by using voltage waveforms in the power train. Active drive is the technique usually used in synchronous Buck converters as shown in Figure 26.1. It can also be used in other topologies, such as the forward converter shown in Figure 26.22. The advantage of this type of drive is that the gate drive voltage is predetermined and fixed for all conditions. The disadvantage is increased circuit complexity (accurate timing of the gate drive waveforms is important) and the gate drive power loss issue discussed above. Some topologies lend themselves to driving the rectifier MOSFETs using voltage waveforms derived from the power circuit, directly, without logic controlled drive ICs. A forward converter example is shown in Figure 26.23. In this example the gate drive is provided by the power transformer winding. This is generally called a "self-driven" gate drive circuit. The advantages are simplicity, automatic

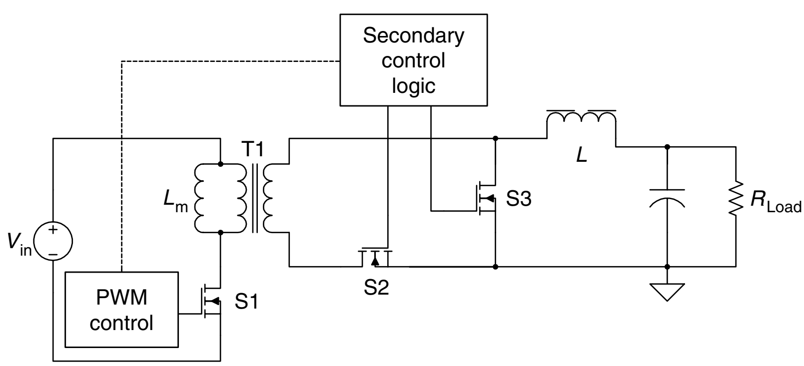


FIGURE 26.22 Forward converter with synchronous rectifiers.

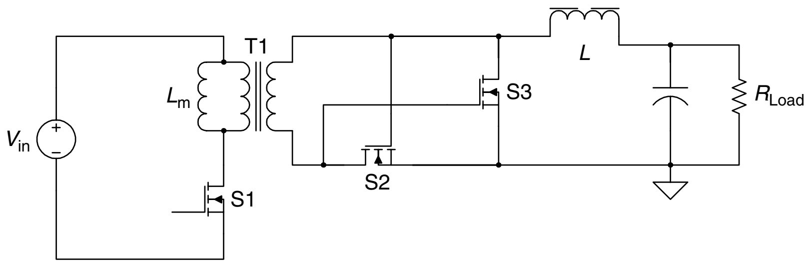


FIGURE 26.23 Forward converter with self-driven synchronous rectifiers.

synchronization with the primary power switch, and recovery of some of the gate drive power that would be lost in a logic driven circuit. The major disadvantages of the self-driven technique are a gate drive voltage that varies with input voltage variations, possible damage to the synchronous MOSFETs under fault conditions, and loss of proper gate drive during light load conditions (discontinuous mode operation). Further discussion of self-driven synchronous rectifier trade-offs is beyond the scope of this chapter, but many technical papers on the subject are available.

Another drive-related issue with synchronous rectifiers, which can be very important, is their behavior at light load conditions. Since a MOSFET device is a bidirectional switch, if the synchronous rectifier is driven complementary to the high-side switch (in other words, when the high-side switch is off the synchronous rectifier switch is on and vice versa, as in Figure 26.1) the peak-to-peak inductor ripple current always remains continuous; there is no discontinuous state. This can have an advantage in noise-sensitive RF circuits since the ripple voltage remains constant independent of load and the switching frequency also stays constant. The converter never enters a pulse-skipping mode that can cause problems in noise-sensitive circuits since the switching frequency and ripple frequency become more chaotic. Since the MOSFET is bidirectional, a synchronous rectifier circuit operated in this complementary mode can also return energy to the source—it operates as a two quadrant converter that can source or sink current. However, if the source powering the synchronous converter cannot sink current, the input voltage will rise as energy is returned from the load. This could potentially over voltage the input capacitors or the MOSFET switches if significant sinking of current occurs under some load condition. Another advantage of this mode of operation is that the transient load performance is consistent from no load to full load—this is not the case for converters that operate in both continuous and discontinuous modes of operation. The disadvantage of this operational mode is that at very light loads energy is still being stored in the inductor and returned to the source. This shuttling of energy into and out of the inductor results in significant conduction losses even at the light load condition. In addition, gate drive power remains the same as what is needed at full load. This is a major disadvantage in battery operated circuits that operate at light loads much of the time and can significantly reduce battery life. Allowing the converter to transition into discontinuous mode (inductor current drops to zero and stays there until the high-side switch is turned back on) usually results in higher efficiency at light loads. Implementation of this mode requires more complex control circuitry when using synchronous rectifiers, whereas, in a diode rectifier circuit this occurs naturally. In a synchronous rectifier circuit, the decrease of rectifier switch current to zero must be sensed and this information used to decide when to turn off the synchronous rectifier.

A third method for controlling the synchronous rectifier is to use some type of pulse-skipping control mode at light loads. At very light loads the converter drive is shut off until a large enough decrease in output voltage is detected, then the switches operate for some number of pulses until the voltage is restored to some fixed level. Switch drive is then shut off until output voltage again drops to a low enough level to initiate another burst of switching. In this way the converter is basically operated in a hysteretic mode with short periods of operation. Different controller IC manufacturers use different variations of this approach. It can yield higher light load efficiencies, since gate drive power only needs to be supplied during the short period switch of operation. The disadvantage is a less controlled noise spectrum and, generally, increased peak-to-peak output voltage ripple compared to continuous mode operation.

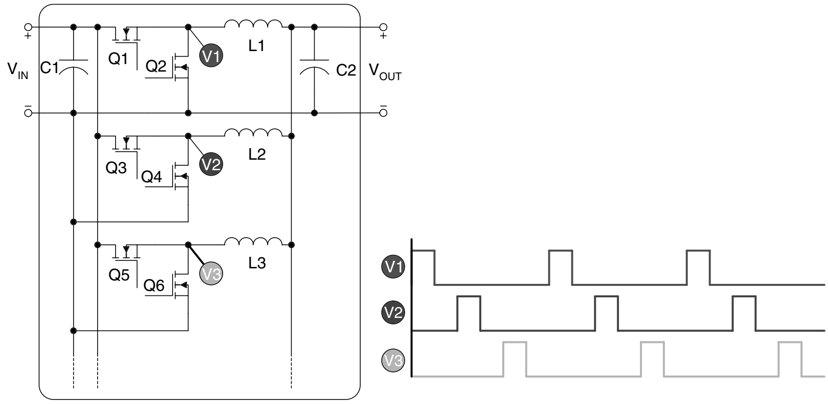


FIGURE 26.24 Multiphase Buck converter.

26.3.2 Multiphase

In recent years, the multiphase Buck converter has become the topology of choice for delivering power to microprocessors. Its success is most likely attributed to several things among them: simplicity, reduced RMS ripple currents in input and output capacitors, thermal performance, and cost. This topology uses multiple Buck converter stages operating in parallel, each phase shifted by $360°/N_\phi$, where N_ϕ is the number of Buck stages (commonly referred to as phases). Figure 26.24 illustrates the basic circuit elements and waveforms in a three phase design.

A direct result of the phase shifting is reduced ripple current in the input and output capacitors (C1 and C2). The amount of ripple current reduction is a function of the input voltage, duty cycle and number of phases. Under certain conditions the ripple current in the output capacitor theoretically reduces to 0 A. It also has the added advantage of spreading the power dissipation out among more components, thus allowing for cooler operation. For example, each inductor in the above schematic only carries one-third of the load current, as does each MOSFET pair. This also has the added advantage of allowing the design to be optimized to use components more commercially available. To the computer industry this means reduced cost. At the time of this article 30 V MOSFETs capable of carrying 20–40 A of load current represent a cost-effective design point. The concept of multiphase can also be extended to many of the other power conversion topologies. When done the advantages/disadvantages may not be the same as they are for the Buck converter.

26.4 Conclusion

Simply put a DC-to-DC power supply takes one DC voltage and puts out a different one. However, the process of getting there can be much more complex. Consideration needs to be given to the system and environment in which the supply is expected to operate (e.g., input voltage, EMI/EMC, energy hold up, output ripple, efficiency, thermal management, battery life, temperature, health monitoring, stability, transient response, etc.). These types of issues all impact the type of supply needed (e.g., linear, charge pump, SMPS) and how that supply needs to be configured for a given application.

As stated earlier, the trend for power seems to be toward lower output voltages at higher load currents. This trend is going to demand that power supply efficiency be maximized. This is always important, however, when the required output voltage drops, the detrimental effect of circuit parasitics (e.g., resistance, diode forward voltage drops) on efficiency becomes larger. This coupled with the fact that the load currents are not dropping, but rising, creates a greater thermal load on the system. In the case of microprocessors, this will likely result in dense hybridized power supplies that will be collocated with the processor.

References

1. Pressman, A. I., *Switching Power Supply Design*, McGraw-Hill, New York, 1991, pp. 248–246.
2. MIL-STD-461C, Electromagnetic Emission and Susceptibility Requirements for the Control of Electromagnetic Interference, Department of Defense, The United States of America, August 4, 1986, Part 2, 13.
3. Mitchell, D. M., *Switching Regulator Analysis*, McGraw-Hill, New York, 1988, pp. 51–73.
4. Sun, J., Mitchell, D. M., and Jenkins, D. E., Delay effects in averaged modeling of PWM converters, Power Electronic Specialists Conference, 1999. PESC99. 30th Annual IEEE, 1999.
5. Modeling, Analysis and Compensation of the Current-Mode Converter, Unitrode Application Note, U97, 1997, Section 3, pp. 43–48.
6. Ridley, R. B., A New, Continuous-Time Model for Current-Mode Control, *IEEE Trans. Power Electr.*, 6, 2, 271–280, 1991.

27

Low Voltage/Low Power Microwave Electronics

Mike Golio
HVVi Semiconductor

27.1 Introduction

The development and manufacture of low voltage/low power radio frequency (RF) and microwave electronics has been the subject of intense focus over the past few decades. The development of active circuits that consume lower power at lower voltage levels is consistent with the general trends of semiconductor scaling. From the simple relationship

$$E = \frac{V}{d} \tag{27.1}$$

where E is the electric field, V is the voltage, and d is the separation distance, it can be seen that smaller devices (reduced d) require less voltage to achieve the same electric field levels. Since to first order it is electric field in a material that dominates electron transport properties, electric field levels must be preserved when devices are scaled. Therefore, for each reduction in critical dimension (gate length or base width) reductions in operating voltage have been realized.

Because semiconductor physics implies an inverse relationship between device size and required voltage level, and because transistor scaling has been a major driver for integrated circuit (IC) technology since the earliest transistors were introduced, the trend toward reduced power supply voltage has been continual since the beginning of the semiconductor industry. The explosive growth of portable wireless products, however, has brought a sense of urgency and technical focus to this issue. Portable product requirements have forced technologists to consider battery drain and methods to extend the battery lifetime.

Since the total global volume requirements for digital ICs far exceeds the requirements for RF and microwave ICs, digital applications have been the primary drivers for most IC device technologies. The continuing advances of complementary metal-oxide semiconductor (CMOS) technology in particular have supported the low voltage/low power trend. Bipolar, BiCMOS, and more recently SiGe heterojunction bipolar transistor (HBT) developments have similarly supported this trend. For digital circuitry, the low voltage/low power trend is essentially without negative consequences. Each reduction in device size and power supply requirement has led to significant performance improvements. The reduced device size leads to smaller, lighter, cheaper, and faster circuits while the reduced voltage leads to reduced power consumption and improved reliability. The combination of these advantages provides the potential for higher levels of integration, which lead to the use of more complex and sophisticated architectures and systems.

For RF and microwave circuitry, however, the push toward lower voltage and lower power consumption is counter to many other radio requirements and presents challenges not faced by digital circuit designers. Although smaller, lighter, cheaper, and faster are generally desirable qualities for RF as well as digital applications, reduced voltage can produce negative consequences for RF circuit performance. In particular, reduced voltage leads to reduced dynamic range, which can be especially critical to the development of linear power amplifiers (PAs).

An imperative to develop and produce low voltage/low power RF products requires that every aspect of radio component development be considered. The ultimate success of a low voltage/low power, RF product strategy is affected by all of the following issues:

- Materials Technology (GaAs, Si, epitaxiy, implant, heterojuctions)
- Device Technology (BJT, HBT, MESFET, MOSFET, HEMT, MEMS)
- Circuit Technology (matching and topology challenges for low voltage)
- Radio System Design Issues (impact of system and radio architecture on performance)

27.2 Motivations for Reduced Voltage

Commercial handheld products are clearly a significant driver for the low voltage/low power imperative. Reduced power consumption translates into longer battery lifetime and/or smaller, lighter products through reduction of battery size. Consumers value both of these attributes highly which has led to the widespread acceptance of portable pagers, cell phones, and wireless personal area network products. An additional collateral benefit to reduced voltage operation is improved product reliability.

For vehicle-based products, the low power imperative is less compelling since battery size and lifetime issues are not as critical to these applications. Smaller and lighter electronics, however, are still valued in these applications. Power supplies represent a fixed payload that must be transported throughout the lifetime of the vehicle. In the case of jet aircraft, a single additional pound of payload represents tens of thousands of dollars of jet fuel over the life of the aircraft. A reduction in the size and weight of power supplies also represents payload that can be displaced with additional safety, communication, or navigation equipment. As in the case of handheld equipment, reduced power consumption will contribute to improved product reliability.

Advantages of low voltage/low power RF circuitry are also significant for military applications. Supply line requirements during a military operation are often a determining factor in the outcome of battle. Transport vehicles are limited in number and capacity, so extended battery lifetime translates directly into increased battlefield transport capacity available for other supplies. When battery resupply requirements are reduced, other valuable resources, such as additional lifesaving materials, can be transported in their place. The advantage is also directly felt by the foot soldier that is able to achieve greater mobility and/or additional electronic functionality (communication or navigation equipment) without being weighed down with batteries.

As mentioned in the introduction, low voltage/low power trends have been almost without negative impact for most digital applications. Most portable radio products are composed of both digital ICs and RF parts, but manufacturers would prefer to run all of the electronics from a single power supply. RF and microwave circuitry has had to follow the trend to low voltage—even when that trend brings on increasing challenges. This requirement to conform is exacerbated by the fact that the volume of digital circuitry required from semiconductor factories is far higher than the volume of RF circuitry. Applications that would benefit from high voltage operation comprise a nearly insignificant portion of the total semiconductor market; so they receive little attention from most semiconductor manufacturers. The RF designer often has little choice but to use a low voltage IC process.

27.3 Semiconductor Materials Technology

Low voltage/low power operation requires improved efficiency, low parasitic resistance, low parasitic capacitance, and precise control of on-voltage. The selection of the semiconductor material impacts all of these characteristics.

Improvements in efficiency and reductions in parasitic resistance can be achieved by using materials that exhibit increased carrier mobility and velocity. Figure 27.1 plots the carrier velocity in bulk semiconductor material as a function of the applied electric field for several commonly used semiconductors. The carrier (electron or hole) mobility is given by

$$\mu_0 = \frac{v}{E} \quad \text{for small } E \tag{27.2}$$

where v is the carrier velocity in the material and E is the electric field. The figure illustrates that III–V semiconductor materials such as GaAs, GaInAs, and InP exhibit superior electron velocity and mobility characteristics relative to silicon. For this reason, devices fabricated in III–V materials with identical critical dimensions are expected to exhibit some advantages in achieving high efficiency and low parasitic resistance.

Low parasitic capacitance is also a desirable characteristic to ensure low loss and minimal battery drain. Because GaAs and InP offer semi-insulating substrates, they exhibit performance advantages over the much more conductive, and hence poorer performing silicon substrates. Advanced materials strategies such as silicon-on-insulator offer potential advantages even over III–V materials, but at increased cost and complexity.

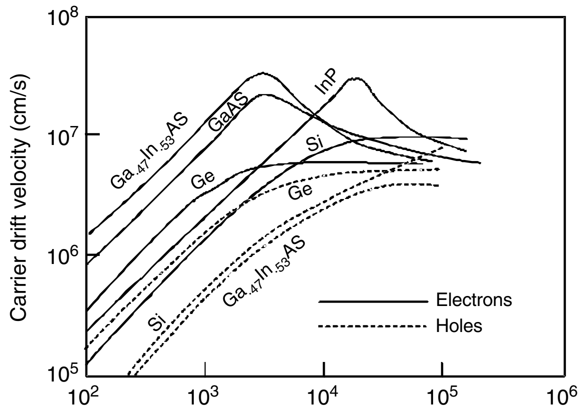


FIGURE 27.1 A plot of the carrier velocity in bulk semiconductor material as a function of the applied electric field for several commonly used semiconductors.

Precise control of on-voltage becomes increasingly important as supply voltages diminish. Low voltage operation means that the total voltage swing that can be realized by the RF signal is reduced. Bipolar transistors offer clear, significant advantages over field-effect transistors (FETs) with respect to this characteristic since bipolar on-voltage characteristic are determined primarily by material bandgaps while FET on-voltage characteristics are determined by material process parameters such as thickness and charge densities. When other considerations indicate that a FET structure is appropriate, however, heterostructure buffer layers improve the on-voltage control of FETs. Within either the FET or bipolar class of semiconductor devices, silicon and SiGe process control tends to be superior to gallium arsenide or other III–V process control.

Heterostructure semiconductor materials offer other potential advantages for low voltage design. In FET structures, heterostructure buffer layers provide greater carrier confinement. This can reduce substrate parasitics. Similarly, in bipolar transistors, heterostructures provide the potential to reduce unwanted minority carrier concentrations in critical portions of the device. In both the FET and bipolar case, improved output conductance and greater linearity results from the use of properly designed heterostructures. Heterostructure technology also offers the potential of realizing high current and high breakdown simultaneously in a device by placing the right material in the right place.

From a purely technical performance perspective, the above discussion argues primarily for the use of III–V heterostructure devices. These arguments are not complete, however. Silicon and SeGe processing is far more mature than III–V processing. This maturity advantage means that more complex structures can be realized in silicon than in III–V materials. Such device structures can often overcome the inherent material advantages of III–V devices. Most commercial wireless products also have requirements for high yield, high volume, and low cost. These requirements can overwhelm the material selection process and they favor mature processes and high volume experience. The silicon high volume manufacturing experience base is far greater than that of any III–V semiconductor facility. Silicon and SiGe digital parts are routinely produced in volumes that dwarf total global III–V semiconductor device requirements. This digital experience is easily leveraged into the production of silicon or SiGe RF and microwave parts—giving silicon materials a clear advantage in these areas.

There is no short cut for the development of experience and maturity with advanced material structures. As demand for diminishing power consumption in RF products continues to grow, however, experience will be gained and the fundamental physical advantages of III–V heterostructure devices is likely to lead to increasing presence in commercial wireless products. The increased volume capability of Si and SiGe factories also carries a negative impact. Product development cycles tend to be much longer in a large silicon facility than in a smaller III–V based operation. In some cases, a III–V based part can be designed and productized before an initial silicon part can be evaluated.

27.4 Semiconductor Device Technology

Device technology decisions influence many of the same performance metrics that are affected by material decisions. Efficiency, parasitic resistance and on-voltage control are all affected by the choice of transistor type. Additional device considerations may depend on the particular circuit application being considered. For an oscillator application, for example, the $1/f$ noise characteristics of the transistor are likely to be critical to achieving the required circuit level performance. These same transistor characteristics may be of little interest for many other circuit applications.

Table 27.1 evaluates several modern transistor technologies against a list of performance characteristics that are of interest for portable PA applications. Not included in the table is standard CMOS or BiCMOS, but even these device technologies are beginning to be explored in RF portable PA applications. Although the RF performance of a single CMOS device is substandard to these other devices in several ways, the cost and integration advantages offered by CMOS devices could eventually overwhelm the other performance drivers for these applications. The performance metrics considered in the table include parasitic loss, ability to easily use a single polarity power supply, power added efficiency, linearity, cost, and maturity.

TABLE 27.1 Semiconductor Devices are Evaluated against Several Important Device Characteristics for Portable Power Amplifier Applications

	MESFET/HEMT	III–V HBT	Si BJT	SiGe HBT	Power MOSFET (LDMOS)
Low voltage related characteristics					
Parasitic loss	Very good	Very good	Moderate	Moderate	Moderate
Single-polarity supply	No[a]	Yes	Yes	Yes	Yes
Power added efficiency	Excellent	Very good	Moderater	Moderate	Good for $f \lesssim 2$ GHz
Linearity	Excellent[b]	Very good	Moderate	Very good	Moderate
Power density	Moderate	Excellent	Very good	Excellent	Moderate
General characteristics					
Cost	Moderate to high	Moderate to high	Low to moderate	Moderate to high	Low to moderate
Maturity	Good	Good	Excellent	Moderate	Very good

[a] e-mode MESFETs and HEMTs do make use of single-polarity supply voltage, but at a cost of more difficult manufacturing and with a more limited dynamic range.
[b] MESFET devices manufactured using epitaxial material (as opposed to ion-implanted) as well as HEMTs exhibit excellent linearity.

As is readily seen from the table, no particular device type excels in all areas and all devices exhibit at least some significant disadvantages. The relative values of each of the metrics are not equal and are not identical for all applications. Meeting linearity requirements, for example, is central to the PA design process for many digitally modulated portable radios, while the linearity requirements for many analog radios are easily achieved and therefore less valued.

Parasitic loss is clearly important to reduce undesirable current drain. The III–V device technologies offer advantages over silicon technologies in this area because of their material advantages.

Single polarity power supplies are desirable since other radio components do not require a negative supply. The generation of a negative voltage required to bias a transistor adds to the parts count, cost and electrical current drain of the radio. HBT and bipolar junction transistor (BJT) parts have a clear advantage over conventional metal-semiconductor field effect transistors (MESFETs) and high electron mobility transistors (HEMTs) in this area. Enhancement mode HEMTs offer a potential solution to this problem, but at a cost to manufacturing ability and dynamic range.

Linearity requirements are typically driven by government regulations to keep wireless products from interfering with each other. The specifications for linearity are not consistent for all radio architectures and are sometimes the subject of erroneous discussions in the technical literature. The linearity of the Gummel plot [$\log(I_c)$ vs. $\log(I_b)$] of a BJT or HBT, for example, is not equivalent or translated into RF linearity. It is also difficult to measure linearity and determine an absolute performance limit for a particular device. RF circuit tuning performed to optimize power, gain, or efficiency is not necessarily optimum for achieving high linearity. MESFETs fabricated using epitaxial material (as opposed to ion-implanted material) and HEMTs exhibit slight advantages in linearity over other devices. HBTs and BJTs have also demonstrated competitive linearity characteristics.

Once power, gain, and linearity specifications are met, efficiency often becomes the key distinguishing figure-of-merit for modern wireless PAs. Improvements in PA efficiency are translated directly into reduced battery drain and longer battery lifetime. HEMTs hold certain advantages in this area while silicon BJTs have not produced efficiency values that are as competitive. HBT parts have also produced competitive efficiencies for most applications.

High power density is desirable since it leads to reduced part size. Smaller devices not only reduce the size of the final product, but also more importantly, reduce the cost of the device since more devices can be processed on a single wafer. HBTs demonstrate the highest power density with adequate breakdown characteristics of any wireless PA devices today. Ion-implanted MESFETs exhibit very low power densities compared to HBTs.

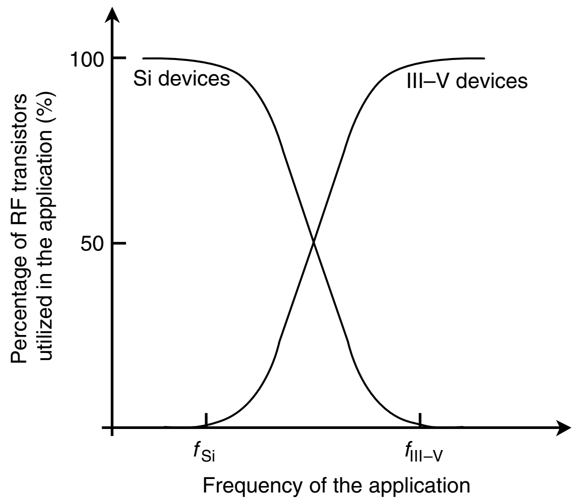


FIGURE 27.2 An illustration of the relationship between frequency of the application and device technology choice.

As in the case of material considerations, the technical characteristics of devices must be weighed against the cost and maturity requirements for wireless product development. Although cost is a primary driver for many radio applications today, it may be difficult to determine fabrication costs of emerging device technologies. For most parts utilized in high volume commercial applications, the cost of fabrication is not closely related to the purchase price for the part. The market supply and demand forces determine purchase price—the price is set as high as the market will bear. If the fabrication costs can be brought below the market price with a comfortable margin, the part will continue to be manufactured. If not, the part will be discontinued. Large semiconductor manufacturing companies may choose to subsidize new technologies for several years before deciding to discontinue them. Although difficult to quantify, cost is critical to the long-term success of a device. Silicon transistors have historically come with a cost advantage over III–V devices. The high volume experience of silicon fabrication facilities and greater maturity of the processes contribute significantly to the low cost advantage.

Examination of Table 27.1 and the discussion of the preceding paragraphs illustrate the difficulty in choosing the optimum device for low voltage/low power applications. Although important, the low voltage characteristics of the part are not the only characteristics that must be considered. No transistor is optimal for all requirements and each device choice carries implied compromises for the final product.

Although not included Table 27.1, the frequency of the application can also be a critical performance characteristic in the selection of device technology. Figure 27.2 illustrates the relationship between frequency of the application and device technology choice. Because of the fundamental material characteristics illustrated in Figure 27.1, silicon technologies will always have lower theoretical maximum operation frequencies than III–V technologies. The higher the frequency of the application, the more likely the optimum device choice will be a III–V transistor over a silicon transistor. Above some frequency (labeled f_{III-V} in Figure 27.2) III–V devices dominate the transistors of choice, with silicon playing no significant role in the RF portion of the product. In contrast, below some frequency (labeled f_{Si} in Figure 27.2) the cost, and maturity advantages of silicon provides little opportunity for III–V devices to compete. In the transition spectrum between f_{Si} and f_{III-V} the device technology is not an either/or choice, but rather silicon and III–V devices coexist. Although silicon devices are capable of operating above frequency f_{Si}, this operation is often gained at the expense of current drain. As frequency is increased above f_{Si} in the transition spectrum, efficiency advantages of gallium arsenide and other III–V devices provide competitive opportunities for these parts. The critical frequencies, f_{Si} and f_{III-V} are not static frequency values and are also dependent on required power levels. These frequencies are continually being moved upward

TABLE 27.2 Semiconductor Devices Are Evaluated against Several Important Device Characteristics for Radio Integration Applications

	MESFET/HEMT	III–V HBT	Si BJT	SiGe HBT	CMOS
Low voltage related characteristics					
Single-polarity supply	No[a]	Yes	Yes	Yes	Yes
Turn-on voltage control	Moderate	Very good	Excellent	Excellent	Good
General characteristics					
Cost	Moderate to high	Moderate to high	Low to moderate	Moderate to high	Low
Maturity	Good	Good	Excellent	Moderate	Excellent
RF integration capability[b]	Very good	Very good	Excellent for $f < 3$ GHz	Excellent for $f < 5$ GHz	Excellent for $f < 2$ GHz
Noise figure	Excellent	Very good	Good	Very good	Moderate
Phase noise	Poor	Good	Excellent	Excellent	Good

[a] e-mode MESFETs and HEMTs do make use of single-polarity supply voltage, but at a cost of more difficult manufacturing and with a more limited dynamic range.
[b] Comparing Silicon RF integration capability to III–V RF integration capability is difficult. Silicon process maturity allows for far higher levels of integration (in terms of numbers of transistors), but becomes of limited value above a few GHz because of the parasitics of the Si substrate.

by the advances of silicon technologies—primarily by decreasing critical device dimensions. Silicon and SiGe operating frequencies are also being extended through the creation of more complex structures that reduce device parasitics.

The process of choosing device technology is compounded further if higher levels of integration are implemented. Table 27.2 is a companion to Table 27.1. Again, several semiconductor device technologies are ranked in terms of desirable device characteristics. In contrast to Table 27.1, Table 27.2 considers characteristics of interest for integrated radio front-end applications. Although some characteristics are common to both tables, many are not. The evaluation of which technology is superior is likely to change when Table 27.2 characteristics are used to make a device technology choice instead of those listed in Table 27.1.

One class of RF application not considered in Tables 27.1 or 27.2 is high power PAs. These applications include base station PAs and usually require amplifiers to produce several dozen to hundreds of Watts of output RF power. High power PAs have not seen pressure to reduce supply voltage. For these applications thermal management issues and high breakdown voltage are often the driving factors that determine device choice. Lateral double diffuse MOS (LDMOS) transistors operating at 28 V have dominated base station applications for over a decade and are being challenged today by wide bandgap semiconductor devices and devices operating at voltages as high as 50 V.

Also not included in the tables above are characteristics of micro-electro-mechanical system (MEMS) devices. MEMS devices offer potential to serve as high performance switches in portable radio devices. Traditionally, these devices have operated only with very high voltages (dozens of volts). Recent research, however, is exploring low voltage MEMS devices for switching applications in portable radios.

27.5 Circuit Design

RF circuit performance is not generally improved by the reduction of voltage or current. Virtually all types of circuits exhibit degradation in performance with reduced DC power. As an example, when either current or voltage is reduced, a typical RF amplifier circuit will exhibit degradation of gain, linearity, and dynamic range.

If RF power levels are to be maintained, PA load lines must be altered as supply voltage is reduced. Figure 27.3 illustrates the issue of load line and reduced voltage. The load line labeled "a" represents an optimum, idealized, high-efficiency load-line for 6 V power supply operation. The slope of the load line

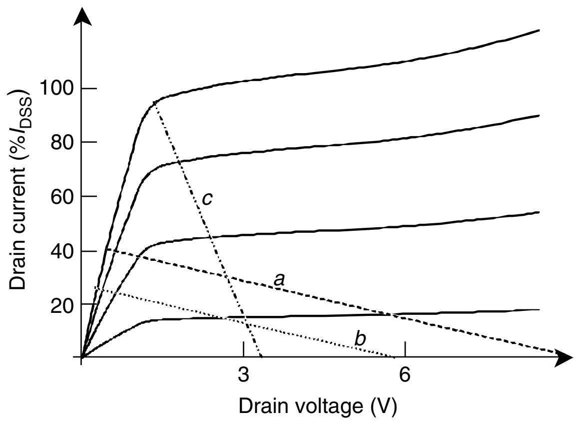


FIGURE 27.3 An illustration of the effect that lowering supply voltage has on PA load line. Load line *a* represents an optimum high efficiency load line for 6 V supply voltage. Load line *b* results from reducing the supply voltage to 3 V without modifying the load line slope. The result of using load line *b* is a reduction in output power by more than 3 dB. Load line *c* represents a 3 V load line where output power level is maintained. The result of using load line *c* is reduced efficiency from the load line *a* case.

is determined primarily by the matching characteristics between the device and the load. As voltage is reduced from 6 to 3 volts, preservation of the slope of the load line would imply a reduction in power by approximately 3 dB (load line labeled "b"). Power is estimated to first order as the area underneath the load line and contained by the current–voltage plot. If power is to be maintained, the load line for the 3 V supply case must be made steeper. This is indicated by load line "c." With a steeper load line, the average current along the load line (which represents DC current drain) is increased, while the intersection of the load line and the I–V curve is raised in voltage. When RF power levels are held constant, moving from load-line "a" to load-line "c" will tend to have the undesirable effect of reducing the amplifier efficiency.

To compensate for this problem, low voltage devices are designed to be larger than their higher voltage counterparts. Increased periphery devices provide higher peak currents thereby maintaining power level with decreased voltage. Larger devices, however, exhibit lower impedance that must be matched. Thus, low voltage PA parts have high transformation ratios. Typical 1–4 W PA parts for commercial portable wireless products exhibit output impedance of less than a few ohms. High-Q matching elements provide some advantage in achieving required transformations but come at high cost and reduced integration.

As designers struggle to maintain power, gain, and efficiency with diminishing supply voltage, linearity must also be maintained. This requirement brings increasing importance to the use of linearization circuitry, feed-forward, and predistortion schemes. Active bias control is also a valuable circuit design technique for many applications. Bias schemes that cause the transistor bias to rise with increased incident RF power help to provide needed maximum power while maintaining efficiency when lower power levels are needed.

Even for low power RF applications, bias control is critical when limited power supply voltage is available. The problem is compounded by device-to-device variability in current–voltage characteristics. Device stacking opportunities are limited and current control bias schemes, as opposed to voltage control schemes, are often preferred.

27.6 Radio and System Architecture

The ongoing communication technology revolution continues to demand greater bit transfer rates and higher bandwidths from wireless products. These demands lead to greater required linearity and higher frequencies of operation from the RF portion of these products. Broader bandwidth, higher frequency operation contributes to increases in battery drain creating a greater challenge to reduce operating voltage and power consumption.

Providing access to significantly increased information content also forces requirements for greater functionality causing a need for greater circuit complexity. Increasing functionality of the product is viewed as a potential method for establishing competitive advantage in an otherwise regulated environment. Since each new function requires power, this trend also works against the low power consumption imperative.

Other radio architecture specifications, such as the use of digital modulation, eases low voltage/low power design issues. Improved signal-to-noise advantages offered by digital modulation results in reduced output power requirements. Digital modulation schemes also provide the potential for pulsed (vs. CW) operation, reducing power consumption even more.

Some proposed and emerging wireless systems would have a dramatic effect on the war on power consumption—changing the requirements (and the technologies of choice) entirely.

Micro-cell telephone systems and small area Wireless Personal Area Networks (see *RF and Microwave Applications and Systems*) will dramatically reduce the battery requirements for these wireless systems. A smaller geographic cell size reduces the required RF output power of both subscriber and base station amplifiers. The required subscriber transmission power is reduced since the total distance that information must be transmitted is reduced. The base station power is also reduced for the same reason and because the total number of users that will be within the smaller cell is reduced. These reductions in required transmission power translate into dramatically reduced power consumption requirements. The power consumption of the RF portion of the radio becomes insignificant compared to the rest of the electronics. The device and material selection process can also become inconsequential. Although from a consumer point of view, this solution offers significant advantage with no clear down-side, the micro-cell infrastructure installation represents a significant investment for the service provider. The actual complexity of the network is significantly greater for micro-cell systems.

In contrast, Satellite Cellular systems (see *RF and Microwave Applications and Systems*) present significant challenges due to the high instantaneous output power requirements of the PA and low minimum noise figure of the front end LNA.

Although the choice of semiconductor material, type of transistor and circuit topology all affect low voltage design, the radio architecture and system concept can dominate important low voltage performance specifications such as required battery size and battery lifetime.

27.7 Limits to Reductions in Voltage

There are some implications to the reduction of battery size and weight and that will impose serious constraints on the minimum useful operating voltage for portable RF products.

As discussed previously in this section, a primary driver for reduced voltage is the successful evolution of low voltage digital IC technologies. The advantage of running digital and RF circuitry from a single power supply is clear. Issues that limit DC power requirements for RF circuits, however, are fundamentally different from those that limit DC power requirements for associated digital circuits. Digital circuitry is required to store and analyze information that is encoded in a binary manner. This can be accomplished theoretically by the presence or absence of a small charge (even a single electron). Although practical considerations make a single electron memory improbable and movement of even one electron into and out of storage still requires energy, it is clear that binary data can be manipulated with extremely small amounts of energy. The digital system requirements do not impose arbitrary power requirements on the strength of the digital signal.

In contrast, minimum RF power requirements are imposed by the radio system design. The RF signal must be transmitted and received across a specified distance or range of distances. Because power is lost in the radiation process, RF circuits must be able to handle power levels that are determined by the propagation media and transmitter-to-receiver separation. If 1 W of transmission power is required for

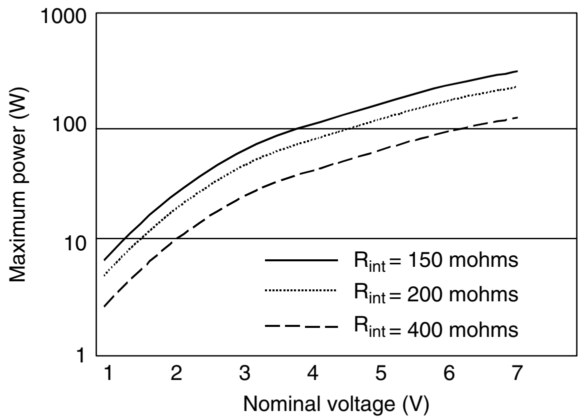


FIGURE 27.4 The maximum power capability for a battery as a function of nominal voltage and internal resistance. Maximum power for this figure is defined as the nominal voltage times the short circuit current.

the remote receiver to successfully detect the signal, then reductions in RF power below this level will cause the radio to fail. Unlike the digital situation where any discernable bit is as good as any other bit, the minimum RF transmission power must be maintained. For RF circuitry, increased efficiency and/or increased current must accompany a reduction in voltage.

Most portable units operate at efficiency levels within 2–10% of the theoretical limits. The implication of near theoretical efficiency is that voltage reductions necessarily involve increased current requirements. If voltage is halved, current must be doubled.

As battery current requirements are increased, the internal resistance becomes a limiting factor in the total power the battery is capable of delivering. For most applications, batteries have been considered ideal voltage sources, but as low voltage electronics trends continue, this assumption will begin to break down.

Figure 27.4 presents the maximum power capability for a battery as a function of nominal voltage and internal resistance. Maximum power for this figure is defined quite optimistically as the nominal voltage times the short circuit current. Although internal resistance is a function of the chemistry as well as the number and size of the battery cells, the values used in Figure 27.4 are typical for nickel-metal-hydride and Li-ion batteries in common use for portable wireless products today. Further, the trends plotted in the figure will hold for all batteries, regardless of battery chemistry or size. It is clear from the figure that as battery voltages are reduced below approximately 2 V, the maximum power available from them is reduced to levels on the order of that required from a portable cellular PA alone. When efficiency and other circuit requirements are accounted for, battery power levels are inadequate to support the product.

27.8 Summary

Reductions in voltage and power consumption are important to the development of portable wireless products that are smaller, lighter and have long battery lifetimes. Low voltage supply requirements are also being driven by a need to make RF ICs compatible with digital ICs. Careful consideration and improvement in the choice of semiconductor material, device type, and circuit topology can help achieve reductions in voltage. System design and choice of radio architecture also has impact on low voltage/low power electronics. These latter issues have potentially the largest impact on low voltage requirements. The ultimate limit to decreasing battery voltage, however, will be determined by transmitter power requirements and achievable internal resistance of batteries.

The discussion in this section has focused on commercial wireless products. These same trends, however, also describe military, satellite, or avionics communications products—both today and over the past several decades. All of these systems are being driven by the demand to move increasing amounts of information between points via a wireless system.

Bibliography

T. Apel, E. Creviston, S. Ludvik, L. Quist, B. Tuch, "A GaAs MMIC transceiver for 2.45 GHz wireless commercial products," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 15–18, 1994.

W. Baumberger, "A single chip rejecting downconverter for the 2.44 GHz band," *GaAs IC Symposium Digest*, pp. 37–40, 1993.

P. Briere, E. Perea, O. de Saint Leger, "System and technology aspects of future digital cellular phones in Europe," *GaAs IC Symposium Digest*, pp. 33–36, 1991.

F. Carrara, A. Castorina, A. Scuderi, G. Palmisano, "High performance silicon bipolar power amplifier for 1.8 GHz applications," *IEEE MTT-S International Microwave Symposium Digest, 2002*, vol. 2, pp. 1015–1018, June 2002.

C. Caux, P. Gamand, M. Pertus, "A cost effective true European DBS low noise converter," *GaAs IC Symposium Digest*, pp. 143–145, 1991.

K. R. Cioffi, "Monolithic L-band amplifiers operating at milliWatt and sub-milliWatt DC power consumption," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 9–12, 1992.

R. Darling, "Subthreshold conduction in uniformly doped epitaxial GaAs MESFETs," *IEEE Transactions on Electron Devices*, vol. 36, pp. 1264–1273, July 1989.

S. Dietsche, C. Duvanaud, G. Pataut, J. Obregon, "Design of high power-added efficiency FET amplifiers operating with very low drain bias voltages for use in mobile telephones at 1.7 GHz," *Proceedings of the European Microwave Conference*, Madrid, pp. 252–254, 1993.

R. C. Dixon, "Portable communications needs and No-Nos," *GaAs IC Symposium Digest*, pp. 29–32, 1993.

G. P. Donzelli, "GaAs versus silicon technology," *Proceedings of the European Microwave Conference*, Madrid, pp. 110–112, 1993.

J. M. Golio, M. Miller, J. Staudinger, W. B. Beckwith, E. N. Arnold, *Microwave MESFETs and HEMTs*, Norwood, MA: Artech House, 1991.

M. Golio, "Large signal microwave device modeling challenges for high volume commercial communications applications," *Workshop Proceedings of the European Microwave Conference*: *Advanced Microwave Devices, Characterisation and Modelling*, Madrid, pp. 30–33, 1993.

M. Golio, "Characterization, parameter extraction and modeling for high frequency applications," *Proceedings of the European Microwave Conference*, Madrid, pp. 69–72, 1993.

M. Golio, "Active devices for commercial subscriber power amplifier products," *1996 URSI Conference*, Lille, France, August 28–September 5, 1996.

M. Golio, "Low voltage microwave electronics," *Santa Clara Valley IEEE MTT Short Course*, Santa Clara, CA, April 4, 1998.

M. Golio, "Low voltage electronics for portable wireless applications: an industrial perspective," *1998 IEEE MTT-S International Microwave Symposium Digest*, Baltimore, MD, June 7–12, 1998.

J. Griffiths, "3 Volt MMIC power amplifier for wireless phones," *Applied Microwave & Wireless*, vol. 6, pp. 38–53, Spring 1994.

D. Hah, S. Hong, "A low-voltage actuated micromachined microwave switch using torsion springs and leverage," *IEEE Transactions on Microwave Theory and Techniques*, vol. 48, pp. 2540–2545, December 2000.

D. Halchin, M. Golio, C. Farley, "Issues with measurements, design, and modeling of devices for wireless applications," *Wireless Conference*, Sedona, AZ, September 30, 1996.

H. Heaney, F. McGrath, P. O'Sullivan, C. Kermarrec, "Ultra low power low noise amplifiers for wireless communications," *GaAs IC Symposium Digest*, pp. 49–51, 1993.

R. Herman, A. Chao, J. Deichsel, C. Mason, A. McKay, "Highly integrated mixed-signal L-band configurable receiver array," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 23–26, 1994.

A. Herrera, E. Artal, E. Puechberty, D. Masliah, "High efficiency, highly compact L-band power amplifier for DECT applications," *Proceedings of the European Microwave Conference*, Madrid, pp. 155–157, 1993.

M. Hirayama, K. Honjo, M. Obara, N. Yokoyama, "HBT IC Technologies and applications in Japan," *GaAs IC Symposium Digest*, pp. 3–6, 1991.

Y. Imai, M. Tokumitsu, A. Minakawa, T. Sugeta, M. Aikawa, "Very low-current and small-size GaAs MMICs for L-band front-end applications," *GaAs IC Symposium Digest*, pp. 71–74, 1989.

Y. Imai, M. Tokumitsu, A. Minakawa, "Design and performance of low-current GaAs MMICs for L-band front-end applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. 39, pp. 209–214, February 1991.

K. Inosako, K. Matsunaga, Y. Okamoto, M. Kuzuhara, "Highly efficient double-doped heterojunction FET's for battery-operated portable power applications," *IEEE Electron Device Letters*, vol. 15, pp. 248–250, July 1994.

O. Ishikawa, Y. Ota, M. Maeda, A. Tezuka, H. Sakai, T. Katoh, J. Itoh, Y. Mori, M. Sagawa, M. Inada, "Advanced technologies of low-power GaAs ICs and power modules for cellular telephones," *GaAs IC Symposium Digest*, pp. 131–134, 1992.

I. Kanno, E. Hironobu, T. Suzuki, H. Kotera, "Piezoelectric Micro-actuators for RF-MEMS Switches," *Pacific Rim Conference on Lasers and electro-Optics, 2005. CLEO/Pacific Rim 2005*, pp. 1380–1381, 30 July–02 August, 2005.

K. Kanazawa, M. Kazumura, S. Nambu, G. Kano, I. Teramoto, "A GaAs double-balanced dual-gate FET mixer IC for UHF receiver front-end applications," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-33, pp. 1548–1553, December 1985.

B. Khabbaz, A. Douglas, J. DeAngelis, L. Hongsmatip, V. Pelliccia, W. Fahey, G. Dawe, "A high performance 2.4 GHz transceiver chip-set for high volume commercial applications," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 11–14, 1994.

C. Kermarrec, G. Dawe, T. Tewksbury, B. Meyerson, D. Harame, M. Gilbert, "SiGe technology: applications to wireless digital communications," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 1–4, 1994.

M. Kim, A. Oki, J. Camou, P. Chow, B. Nelson, D. Smith, J. Canyon, C. Yang, R. Dixit, B. Allen, "12–40 GHz low harmonic distortion and phase noise performance of GaAs heterojnuction bipolar transistors," *GaAs IC Symposium Digest*, pp. 117–120, 1988.

Y. Kitahara, "Personal mobile communications in microwave band," *Proceedings of the European Microwave Conference*, Madrid, pp. 94–95, 1993.

K. W. Kobayashi, R. Esfandiari, M. Hafizi, D. Streit, A. Oki, M. Kim, "GaAs HBT wideband and low power consumption amplifiers to 24 GHz," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 85–88, 1991.

T. Kunihisa, T. Yooyama, H. Fujimoto, K. Ishida, H. Takehara, O. Ishikawa, "High efficiency, low adjacent channel leakage GaAs power MMIC for digital cordless telephone," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 55–58, 1994.

L. Lu, Z. Wang, "Analysis and design of a low-voltage RF CMOS mixer," *IEEE Transaction on Circuits and Systems II: Express Briefs*, vol. 53, pp. 212–216, March 2006.

W. Liu, T. Kim, P. Ikalainen, A. Khatibzadeh, "High linearity power X-band GaInP/GaAs heterojunction bipolar transistor," *IEEE Electron Device Letters*, vol. 15, pp. 190–192, June 1994.

S. Maas, B. Nelson, D. Tait, "Intermodulation in heterojunction bipolar transistors," *IEEE Microwave Theory and Techniques Symposium Digest*, pp. 91–93, 1991.

M. Maeda, M. Nishijima, H. Takegara, C. Adachi, H. Fujimoto, Y. Ota, O. Ishikawa, "A 3.5 V, 1.3 W GaAs power multi chip IC for cellular phone," *GaAs IC Symposium Digest*, pp. 53–56, 1993.

S. Makioka, K. Tateoka, M. Yuri, N. Yoshikawa, K. Kanazawa, "A high efficiency GaAs MCM power amplifier for 1.9 GHz digital cordless telephones," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 51–54, 1994.

R. Meyer, W. Mack, "A 1-GHz BiCMOS RF front-end IC," *IEEE Journal of Solid-State Circuits*, vol. 29, pp. 350–355, March 1994.

J. Moniz, B. Maoz, "Improving the dynamic range of Si MMIC gilbert cell mixers for homodyne receivers," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 103–106, 1994.

V. Nair, "Low current enhancement mode MMICs for portable communication applications," *GaAs IC Symposium Digest*, pp. 67–70, 1989.

V. Nair, R. Vaitkus, D. Scheitlin, J. Kline, H. Swanson, "Low current GaAs integrated down converter for portable communication applications," *GaAs IC Symposium Digest*, pp. 41–44, 1993.

M. Nakatsugawa, Y. Yamaguchi, M. Muraguchi, "An L-band ultra low power consumption monolithic low noise amplifier," *GaAs IC Symposium Digest*, pp. 45–48, 1993.

Y. Ota, M. Yanagihara, T. Yokoyama, C. Azuma, M. Maeda, O. Ishikawa, "Highly eficient, very compact GaAs power module for cellular telephone," *IEEE Microwave Theory and Techniques Symposium Digest*, pp. 1571–1520, 1992.

P. O'Sullivan, G. St. Onge, E. Heaney, F. McGrath, C. Kermarrec, "High performance integrated PA, T/R Switch for 1.9 GHz personal communications handsets," *GaAs IC Symposium Digest*, pp. 33–35, 1993.

J. Pedro, J. Perez, "A novel non-linear GaAs FET model for intermodulation analysis in general purpose harmonic-balance simulators," *Proceedings of the European Microwave Conference*, Madrid, pp. 714–716, 1993.

D. Peroulis, S. P. Pacheco, K. Sarabandi, L. P. B. Katehi, "Electromechanical considerations in developing low-voltage MEMS switches," *IEEE Transactions on Microwave Theory and Techniques,* vol. 51, pp. 259–270, January 2003.

E. Pettenpaul, "New horizons for the microwave business in Europe," *Proceedings of the European Microwave Conference*, Madrid, pp. 1–5, 1993.

E. Pettenpaul, L. Scharf, K.J. Schopf, "Enhanced GaAs device concepts for the new digital mobile communication systems," *Proceedings of the European Microwave Conference*, Madrid, pp. 132–134, 1993.

P. Philippe, M. Pertus, "A 2 GHz enhancement mode GaAs down converter IC for satellite TV tuner," *Proceedings of the European Microwave Conference*, Madrid, pp. 61–64, 1993.

J. O. Plouchart, H. Wang, M. Riet, "HBT monolithic integrated phase locked oscillator for DCS 1800 mobile communications," *Proceedings of the European Microwave Conference*, Madrid, pp. 834–836, 1993.

R. A. Pucel, D. Masse, R. Bera, "Performance of GaAs FET mixers at X-band," *IEEE Transactions on Microwave Theory and Techniques*, vol. MTT-24, pp. 351–360, June 1976.

K. Sakuno, M. Akagi, H. Sato, M. Miyauchi, M. Hasegawa, T. Yoshimasu, S. Hara, "A 3.5 W HBT MMIC power amplifier module for mobile communications," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 63–66, 1994.

N. Shiga, T. Sekiguchi, S. Nakajima, K. Otobe, N. Kuwata, K. Matsuzaki, H. Hayashi, "MMIC family for DBS downconverter with pulse-doped GaAs MESFETs," *GaAs IC Symposium Digest*, pp. 139–142, 1991.

M. Sullivan, "Wireless LANs, status and future," *Applied Microwave & Wireless*, vol. 6, pp. 54–61, Spring 1994.

H. Takeuchi, M. Muraoka, T. Hatakeyama, A. Matsuoka, M. Honjo, S. Miyazaki, K. Tanaka, T. Nakata, "A Si wide-band MMIC amplifier family for L-S band consumer product applications," *IEEE Microwave Theory and Techniques Symposium Digest*, pp. 1283–1284, 1991.

L. Vestling, R. Valtonen, U. Heinle, J. Ankarcrona, J. Olsson, "Improved output conductance for low-voltage microwave LDMOS transistors," *Proceeding of the 30th European Solid-State Device Research Conference, 2000*, pp. 468–471, 11–13 September, 2000.

D. Vook, T. Kamins, G. Burton, P. Vande Voorde, H. Wang, R. Coen, J. Lin, et al., "Double-diffused graded SiGe-Base bipolar transistors," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1013–1018, June 1994.

M. Wang, M. Carriere, P. O'Sullivan, B. Maoz, "A single-chip MMIC transceiver for 2.4 GHz spread spectrum communication," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 19–22, 1994.

D. Whipple, "The CDMA standard," *Applied Microwave & Wireless,* vol. 6, pp. 24–37, Spring 1994.

T. Yamaguchi, T. Archer, R. Johnston, J. Lee, "Process and device optimization of an analog complementary bipolar IC technology with 5.5-GHz fT PNP transistors," *IEEE Transactions on Electron Devices*, vol. 41, pp. 1019–1026, June 1994.

T. Yoshimasu, K. Sakuno, N. Matsumoto, E. Suematsu, T. Tsukao, T. Tomita, "Low current GaAs MMIC family with a miniaturized band-stop filter for Ku-band broadcast satellite applications," *GaAs IC Symposium Digest*, pp. 147–150, 1991.

T. Yoshimasu, N. Tanaba, S. Hara, "An HBT MMIC linear power amplifier for 1.9 GHz personal communications," *IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium Digest*, pp. 59–62, 1994.

"Low voltage silicon bipolar transistors," *Microwave Journal*, vol. 37, pp. 126–128, June 1994.

III

Underlying Physics

28

Maxwell's Equations

Nicholas E. Buris
Motorola Labs

Microwaves and RF is a branch of electrical engineering which deals ultimately with special cases of the physics of electrically charged particles and their interactions via electromagnetic waves. The fundamental branch of science describing the physics of electrically charged particles is electromagnetism. Electromagnetism deals with the electromagnetic force and is based on the concept of electric and magnetic vector fields, $\mathcal{E}(\boldsymbol{r}, t)$ and $\mathcal{H}(\boldsymbol{r}, t)$, respectively. The fields $\mathcal{E}(\boldsymbol{r}, t)$ and $\mathcal{H}(\boldsymbol{r}, t)$ were first introduced to resolve the issues of the "action at a distance" experienced between charges. Maxwell's equations are four coupled partial differential equations describing the electromagnetic field in terms of its sources, the charges and their associated currents (charges in motion). Electromagnetic waves are one special solution of Maxwell's equations that microwave engineering is built upon. In engineering, of course, depending on the technology of interest, we deal with a full range of special circumstances of electromagnetism. At one end of the spectrum are applications such as solid-state devices where electromagnetics is applied to just a few charges, albeit in a phenomenological sense and in conjunction with quantum mechanics. In this realm the forces on individual charges are important. At the other end, we have applications where the wavelength of the electromagnetic waves is much smaller than the dimensions of the problem and electromagnetics is reduced to optics where only simple, plane wave phenomena are at play. In the middle of the spectrum, we deal with structures whose size is comparable to the wavelength and electromagnetics is treated as a rigorous mathematical boundary value problem. The majority of microwave applications is somewhat in the middle of this spectrum with some having connections to either end. When studying time-harmonic events in microwaves, the frequency domain version of Maxwell's equations is very convenient. In the frequency domain, we have developed a number of high level descriptions of electromagnetic phenomena and several specialized disciplines such as circuits, filtering, antennas, and others have been created to efficiently address the engineering problems at hand. This chapter of the handbook will describe Maxwell's equations and their solution in order to establish the connection between the various microwave and *rf* topics and their basic physics, electromagnetism.

28.1 Time Domain Differential Form of Maxwell's Equations

As implied earlier, the fundamental description of the physics involved in the study of microwaves and RF is based on the concept of electric and magnetic vector fields, $\mathcal{E}(\boldsymbol{r}, t)$ and $\mathcal{H}(\boldsymbol{r}, t)$, respectively. According to the well-known Helmholtz' theorem, any vector field can be uniquely specified in terms of its rotation (curl) and divergence components [3]. Maxwell's equations in vacuum essentially define the sources of the curl and divergence of $\mathcal{E}$ and $\mathcal{H}$. In the International System of units (SI) these equations take the following form:

$$\nabla \times \mathcal{E} = \mathcal{J}_{\mathrm{m}} - \mu_{\mathrm{o}} \frac{\partial \mathcal{H}}{\partial \mathrm{t}} \tag{28.1}$$

$$\nabla \cdot \mathcal{E} = \frac{\rho_{\mathrm{e}}}{\varepsilon_{\mathrm{o}}} \tag{28.2}$$

$$\nabla \times \mathcal{H} = \mathcal{J}_{\mathrm{e}} + \varepsilon_{\mathrm{o}} \frac{\partial \mathcal{E}}{\partial \mathrm{t}} \tag{28.3}$$

$$\nabla \cdot \mathcal{H} = \frac{\rho_{\mathrm{m}}}{\mu_{\mathrm{o}}} \tag{28.4}$$

where the constants ε_{o} and μ_{o} are the permitivity and permeability of vacuum, ρ_{e} and $\mathcal{J}_{\mathrm{e}}$ are the electric charge and current densities while ρ_{m} and $\mathcal{J}_{\mathrm{m}}$ are the magnetic charge and current densities.

Equations 28.1–28.4 indicate that, in addition to the charges and the currents, time variation in one field serves as a source to the other. In that sense, in microwaves, dealing with high-frequency harmonic time variations, the electric and magnetic fields are always coupled and we refer to them combined as the electromagnetic field. It is interesting to note that Maxwell developed his equations by abstraction and generalization from a number of experimental laws that had been discovered before him. Up to Maxwell's time, electromagnetism was a collection of interesting experimental and theoretical laws from Coulomb, Gauss, Faraday, and others. Maxwell, in 1864, combined and extended all these into a remarkably complete system of equations thus founding the science of electromagnetism. Maxwell's generalizations helped start and propel work in electromagnetic waves, and also facilitated the introduction of the special theory of relativity. Interestingly, Maxwell's equations are not covariant under a Galilean transformation (observer moving with respect to the environment). However, after the postulates of the special theory of relativity, no modification of any kind was needed to Maxwell's equations. The speed of light, derived from the wave solutions to Maxwell's equations, is a constant for all inertial frames of reference.

As mentioned above, the electromagnetic fields are a conceptual framework, the result of an effort to systematically describe how electrically charged particles move. Maxwell's equations describe the field in terms of its sources but they do not describe how the charges move. The motion of charges is governed by Coulomb's law. The force, $\mathcal{F}$, on an electric charge, q, moving with velocity v inside an electromagnetic field is

$$\mathcal{F} = q\mathcal{E} + \mathrm{q}\mu_{\mathrm{o}} v \times \mathcal{H}$$

A very important assumption made here is that the charge q is a test charge, that is, the charge is small enough that it does not alter the field in which it exists. The problem of the self fields of charges cannot be solved by classical means, but only through quantum electrodynamics [12]. It should be noted that the fields themselves cannot be measured directly. It is through their effects on charged particles that they are experienced. In most microwave applications we do not see Coulomb's force because we seldom deal with just a few particles. Instead, abstractions such as voltage, impedance, and others have been devised to arrive at engineering designs efficiently.

28.2 Some Comments on Maxwell's Equations

Because the divergence of the curl of any vector is identically equal to zero, combining Equations 28.2 and 28.3 results in the current continuity equation (charge conservation), that is,

$$\nabla \cdot \mathcal{J}_{\mathrm{e}} + \frac{\partial \rho_{\mathrm{e}}}{\partial t} = 0 \tag{28.5a}$$

Similarly,

$$\nabla \cdot \mathcal{J}_{\mathrm{m}} + \frac{\partial \rho_{\mathrm{m}}}{\partial t} = 0 \tag{28.5b}$$

This is a peculiar fact, as the continuity equation is somewhat of a statement on the nature of the field sources and, as such, one would not expect it to be an intrinsic property of Maxwell's equations.

Another peculiar property of Maxwell's equations is that they are covariant under a duality transformation (see Section 28.6.2). Critical quantities, quadratic in the fields, remain invariant under such a transformation. Under the proper choice of a duality transformation, Equations 28.1–28.4 can be made to have only electric, or only magnetic sources. Therefore, the question of whether magnetic sources exist is equivalent to whether the ratio of electric to magnetic sources is the same for all charged particles [6]. Again, this is another statement related to the structure of the field sources and it is rather peculiar that it should be contained in Maxwell's equations. In microwave engineering, we frequently encounter complex problems where it is convenient to consider both electric and magnetic sources present. The concept of duality is used extensively to simplify certain problems where apertures are significant parts of the geometry at hand.

28.3 Frequency Domain Differential Form of Maxwell's Equations

It is customary to restrict investigation of Maxwell Equations to the case where the time variations are harmonic, adopting the phasor convention $e^{j\omega t}$ where ω represents the angular frequency. According to this convention, $\mathcal{E}(\boldsymbol{r}, t) = \mathrm{Re}\{\mathbf{E}(\mathbf{r}, \omega)e^{j\omega t}\}$. Similar expressions hold for all scalar and vector quantities. We say that $\mathcal{E}(\boldsymbol{r}, t)$ is the time domain representation of the field while $\mathbf{E}(\mathbf{r}, \omega)$ is the phasor or frequency domain representation. This convention simplifies the mathematics of the partial differential equations with respect to time, reducing time derivatives to simple multiplication by $j\omega$. It should be noted, however, that the product of two harmonic signals in the time domain does not correspond to the product of their phasors.

To effectively treat complicated materials such as dielectrics, ferrites, and others, the electric and magnetic flux density vector fields are introduced, $\mathbf{D}(\mathbf{r}, \omega)$ and $\mathbf{B}(\mathbf{r}, \omega)$, respectively. These fields essentially account for complicated material mechanisms such as losses and memory (dispersion), in a phenomenological way. They represent average field quantities when large quantities of particles are present. In fact, in the Gaussian system of units, **E** and **D** have the same units and so do **H** and **B**. To first-order approximation, for example, $\mathbf{D}(\mathbf{r}, \omega)$, the electric flux density field in a dielectric equals the externally applied electric field plus the field due to the dipole moment, created by the atoms being "stretched" by the external field. Similar arguments could be made for the magnetic field flux density, although, to be correct, quantum mechanical considerations are needed for a more satisfying explanation. Additional discussions

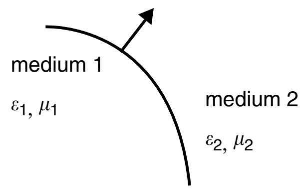


FIGURE 28.1 Geometry for the boundary conditions at the interface between two media.

on materials are made in Chapter 26 of the companion volume "RF and Microwave Passive and Active Technologies". Maxwell's equations in complicated media and in the frequency domain become

$$\nabla \times \mathbf{E} = -\mathbf{J}_{\mathrm{m}} - j\omega\mathbf{B} \tag{28.6a}$$

$$\nabla \cdot \mathbf{D} = \rho_{\mathrm{e}} \tag{28.6b}$$

$$\nabla \times \mathbf{H} = \mathbf{J}_{\mathrm{e}} + j\omega\mathbf{D} \tag{28.6c}$$

$$\nabla \cdot \mathbf{B} = \rho_{\mathrm{m}} \tag{28.6d}$$

with the constitutive relations

$$\mathbf{D} = \overline{\overline{\varepsilon}} \cdot \mathbf{E} \tag{28.7}$$

and

$$\mathbf{B} = \overline{\overline{\mu}} \cdot \mathbf{H} \tag{28.8}$$

and with the continuity equations

$$\nabla \cdot \mathbf{J}_{\mathrm{e}} + j\omega\rho_{\mathrm{e}} = 0 \tag{28.9}$$

$$\nabla \cdot \mathbf{J}_{\mathrm{m}} + j\omega\rho_{\mathrm{m}} = 0 \tag{28.10}$$

In this formulation, $\overline{\overline{\mu}}$ and $\overline{\overline{\varepsilon}}$ are the generalizations of the parameters ε_{o} and μ_{o} and they can be complex, tensorial functions of frequency. There are also some rare materials for which the constitutive relations are even more general. For a simplified, molecular level derivation of ε for dielectrics see the text by Jackson [5].

To study inhomogeneous materials, we need proper boundary conditions to govern the behavior of the fields across the interface between two media. Consider such an interface as depicted in Figure 28.1.

Equations 28.1–28.4 are associated with the following boundary conditions (derivable from Maxwell's equations themselves).

$$\hat{n} \cdot (\mathbf{D}_2 - \mathbf{D}_1) = \rho_{\mathrm{es}} \tag{28.11a}$$

$$\hat{n} \cdot (\mathbf{B}_2 - \mathbf{B}_1) = \rho_{\mathrm{ms}} \tag{28.11b}$$

$$-\hat{n} \times (\mathbf{E}_2 - \mathbf{E}_1) = \mathbf{J}_{\mathrm{ms}} \tag{28.11c}$$

$$\hat{n} \times (\mathbf{H}_2 - \mathbf{H}_1) = \mathbf{J}_{\mathrm{es}} \tag{28.11d}$$

where $\hat{\mathbf{n}}$ is the unit vector normal to the interface pointing from medium 1 to medium 2. The quantities ρ_{es}, $\mathbf{J}_{\mathrm{es}}$, ρ_{ms}, and $\mathbf{J}_{\mathrm{ms}}$, represent the free electric and magnetic charges and current surface densities on the interface, respectively. Note that these are free sources and, as such, only exist on conductors and are equal to zero on the interface between two dielectric media. There are bound (polarization) charges on dielectric interfaces, but they are accounted for by Equation 28.11a, and the constitutive relation Equation 28.7.

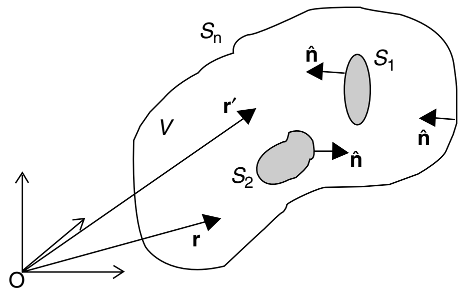


FIGURE 28.2 The volume of interest and its boundary surface consisting of $S_1 + S_2 + \cdots + S_n$.

28.4 General Solution to Maxwell's Equations (The Stratton–Chu Formulation)

One of the most comprehensive solution to Maxwell's equations in a general homogeneous and isotropic domain is given by Stratton and Chu [1] and is also further discussed by Silver [2]. Consider a volume V with a boundary consisting of a collection of closed surfaces, $S_1, S_2, \ldots, S_n$. Next consider the unit vectors, $\hat{\mathbf{n}}$, normal to the boundary surface with direction pointing inside the volume of interest as depicted in Figure 28.2. The volume V is occupied uniformly by a material with dielectric constant ε and magnetic permeability μ. Inside V there exist electric and magnetic charges and current density distributions, $\rho_e, \mathbf{J}_e, \rho_m$, and $\mathbf{J}_m$, respectively.

The electric and magnetic fields at an arbitrary point, $\mathbf{r}$, inside V are then given in terms of the sources and the values of the fields at the boundary by the following equations:

$$\begin{aligned}\mathbf{E}(\mathbf{r}) = &-\int_V \left[j\omega\mu G(\mathbf{r},\mathbf{r}')\mathbf{J}_e(r') + \mathbf{J}_m(\mathbf{r}') \times \nabla' G - \frac{\rho_e(\mathbf{r}')}{\varepsilon}\nabla' G \right] dV' \\ &- \int_{S_1+S_2+\cdots+S_n} [j\omega\mu G(\mathbf{r},\mathbf{r}')(\hat{n}' \times \mathbf{H}(\mathbf{r}')) + (-\hat{n}' \times \mathbf{E}(\mathbf{r}')) \times \nabla' G - (\hat{n}' \cdot \mathbf{E}(\mathbf{r}'))\nabla' G] dS' \end{aligned} \tag{28.12}$$

and

$$\begin{aligned}\mathbf{H}(\mathbf{r}) = &-\int_V \left[j\omega\varepsilon G(\mathbf{r},\mathbf{r}')\mathbf{J}_m(r') - \mathbf{J}_e(\mathbf{r}') \times \nabla' G - \frac{\rho_m(\mathbf{r}')}{\mu}\nabla' G \right] dV' \\ &- \int_{S_1+S_2+\cdots+S_n} [j\omega\mu G(\mathbf{r},\mathbf{r}')(-\hat{n}' \times \mathbf{E}(\mathbf{r}')) - (\hat{n}' \times \mathbf{H}(\mathbf{r}')) \times \nabla' G - (\hat{n}' \cdot \mathbf{H}(\mathbf{r}'))\nabla' G] dS' \end{aligned} \tag{28.13}$$

where

$$G(\mathbf{r},\mathbf{r}') = \frac{e^{-jk|\mathbf{r}-\mathbf{r}'|}}{4\pi|\mathbf{r}-\mathbf{r}'|} \tag{28.14}$$

is called the free space Green's function and

$$k = \sqrt{\omega^2\mu\varepsilon} \tag{28.15}$$

is called the wavenumber with ω representing the angular frequency.

When the surface S_n is at infinity, it can be shown that the fields there satisfy the radiation conditions (**E** and **H** attenuate at least as fast as $1/r$, they are perpendicular to each other and to **r** and their magnitudes are related by the wave impedance, $\sqrt{\mu/\varepsilon}$). Moreover, the continuity equations can be employed to eliminate ρ_e and ρ_m from the field solution. When the volume V is unbounded ($S_n \to$ infinity) and the charges are substituted by their current density expressions, the solution of Maxwell's equations becomes

$$\mathbf{E}(\mathbf{r}) = -\frac{j}{\omega\varepsilon}\int_V [(\mathbf{J}_e(\mathbf{r}')\cdot\nabla')\nabla' + k^2\mathbf{J}_e(\mathbf{r}') - j\omega\varepsilon\mathbf{J}_m(\mathbf{r}')\times\nabla']G(\mathbf{r},\mathbf{r}')dV' \tag{28.16}$$

and

$$\mathbf{H}(\mathbf{r}) = -\frac{j}{\omega\mu}\int_V [(\mathbf{J}_m(\mathbf{r}')\cdot\nabla')\nabla' + k^2\mathbf{J}_m(\mathbf{r}') + j\omega\mu\mathbf{J}_e(\mathbf{r}')\times\nabla']G(\mathbf{r},\mathbf{r}')dV'. \tag{28.17}$$

These forms of the fields are used routinely for the numerical solution to Maxwell's equations, particularly using the method of moments, which is discussed in Chapter 28 of the companion volume "RF and Microwave Circuits, Measurements, and Modeling".

It should be noted here that in the special case where the boundary surfaces $S_1, S_2, \ldots, S_{n-1}$ are perfect electric conductors, the field solution can be expressed by Equations 28.12 and 28.13, or their unbounded counterpart, Equations 28.16 and 28.17, provided one recognizes that the free currents on an electric conductor are $\hat{\mathbf{n}} \times \mathbf{H}$ and the charges are $\hat{\mathbf{n}} \cdot \mathbf{D}$. From the materials point of view a perfect conductor has unlimited capacity to provide free charges that distribute themselves on its surface in such a way as to effectively eliminate their internal fields regardless of the external field they are in. Therefore, since the fields vanish inside, perfect conductors have no energy in them. They simply shape and guide the energy in the space around them. Thus, a coil has no energy inside its metal, it simply stores energy inside and outside its windings. It is this fundamental fact that makes electromagnetic design of especially small, densely populated electronic structures very difficult and prone to electromagnetic interference (EMI) problems.

28.5 Far Field Approximation

At the limit where the observation point is at infinity ($r \to \infty$, far field) it can be shown that Equations 28.16 and 28.17 simplify to

$$\mathbf{E}(\mathbf{r}) = -j\omega\mu\frac{e^{-jkr}}{4\pi r}\int_V \left[\mathbf{J}_e(\mathbf{r}') - \hat{\mathbf{r}}'\mathbf{J}_e(\mathbf{r}') + \sqrt{\frac{\varepsilon}{\mu}}\mathbf{J}_m(\mathbf{r}')\times\hat{\mathbf{r}}'\right]e^{jk\hat{\mathbf{r}}-\mathbf{r}'}dV' \tag{28.18}$$

with

$$\mathbf{H}(\mathbf{r}) = \sqrt{\frac{\varepsilon}{\mu}}\hat{\mathbf{r}}\times\mathbf{E}(\mathbf{r}) \tag{28.19}$$

$\mathbf{E}(\mathbf{r})$ and, consequently, $\mathbf{H}(\mathbf{r})$ in Equations 28.18 and 28.19, have zero radial component and can be found often in the following, more practical, forms.

$$\mathbf{E}_\theta(\mathbf{r}) = -j\omega\mu\frac{e^{-jkr}}{4\pi r}\int_V \left[\mathbf{J}_\theta(\hat{\mathbf{r}}') + \sqrt{\frac{\varepsilon}{\mu}}\mathbf{J}_{m_\phi}(\mathbf{r}')\right]e^{jk\hat{\mathbf{r}}-\mathbf{r}'}dV' \tag{28.20}$$

and

$$\mathbf{E}_\phi(\mathbf{r}) = -j\omega\mu\frac{e^{-jkr}}{4\pi r}\int_V \left[\mathbf{J}_\phi(\mathbf{r}') + \sqrt{\frac{\varepsilon}{\mu}}\mathbf{J}_{m_\theta}(\mathbf{r}')\right]e^{jk\hat{\mathbf{r}}-\mathbf{r}'}dV' \tag{28.21}$$

or, equivalently,

$$\mathbf{E}_\theta(\mathbf{r}) = -\,\mathrm{j}\omega\mu\frac{\mathrm{e}^{-\mathrm{jkr}}}{4\pi\,\mathrm{r}}\int_{\mathrm{V}}[\mathbf{J}_{\mathrm{x}}\cos\theta\cos\phi + \mathbf{J}_{\mathrm{y}}\cos\theta\sin\phi - \mathbf{J}_{\mathrm{z}}\sin\theta$$
$$+\sqrt{\frac{\varepsilon}{\mu}}(-\mathbf{J}_{\mathrm{m_x}}\sin\phi + \mathbf{J}_{\mathrm{m_y}}\cos\phi)]\mathrm{e}^{\mathrm{jk}\hat{\mathbf{r}}\cdot\mathbf{r}'}\mathrm{dV}' \tag{28.22}$$

and

$$\mathbf{E}_\phi(\mathbf{r}) = -\mathrm{j}\omega\mu\frac{\mathrm{e}^{-\mathrm{jkr}}}{4\pi\,\mathrm{r}}\int_{\mathrm{V}}\left[-\mathbf{J}_{\mathrm{x}}\sin\phi + \mathbf{J}_{\mathrm{y}}\cos\phi - \sqrt{\frac{\varepsilon}{\mu}}(\mathbf{J}_{\mathrm{m_x}}\cos\theta\cos\phi + \mathbf{J}_{\mathrm{m_y}}\cos\theta\sin\phi - \mathbf{J}_{\mathrm{m_z}}\sin\theta)\right]\mathrm{e}^{\mathrm{jk}\hat{\mathbf{r}}\cdot\mathbf{r}'}\mathrm{dV}' \tag{28.23}$$

where it also holds that

$$\mathbf{H}_\phi = \sqrt{\frac{\varepsilon}{\mu}}\mathbf{E}_\theta \tag{28.24}$$

and

$$\mathbf{H}_\theta = -\sqrt{\frac{\varepsilon}{\mu}}\mathbf{E}_\phi \tag{28.25}$$

These solutions to Maxwell's equations can also be derived via a potential field formulation. The interested reader can refer to the treatment in Harrington [10] and Balanis [7].

28.6 General Theorems in Electromagnetics

28.6.1 Uniqueness of Solution

It can be shown that the field in the volume V, depicted in Figure 28.2, is uniquely defined by the source distributions in it and the values of the tangential electric field, or tangential magnetic field on all the boundary surfaces, $S_1, S_2, \ldots, S_n$. Uniqueness is also guaranteed when the tangential **E** is specified for part of the boundary while tangential **H** is specified for the remaining part. A detailed discussion of the uniqueness of the solution can be found in Stratton [9].

28.6.2 Duality

As briefly mentioned earlier, Maxwell's equations are covariant under duality transformations. Consider the transformation,

$$\begin{pmatrix}\mathbf{E}\\ \mathbf{H}\end{pmatrix} = \begin{bmatrix}\cos\xi & \sin\xi\\ -\sin\xi & \cos\xi\end{bmatrix}\cdot\begin{pmatrix}\mathbf{E}'\\ \mathbf{H}'\end{pmatrix} \tag{28.26}$$

$$\begin{pmatrix}\mathbf{D}\\ \mathbf{B}\end{pmatrix} = \begin{bmatrix}\cos\xi & \sin\xi\\ -\sin\xi & \cos\xi\end{bmatrix}\cdot\begin{pmatrix}\mathbf{D}'\\ \mathbf{B}'\end{pmatrix} \tag{28.27}$$

$$\begin{pmatrix}\mathbf{J}_{\mathrm{e}}\\ \mathbf{J}_{\mathrm{m}}\end{pmatrix} = \begin{bmatrix}\cos\xi & \sin\xi\\ -\sin\xi & \cos\xi\end{bmatrix}\cdot\begin{pmatrix}\mathbf{J}'_{\mathrm{e}}\\ \mathbf{J}'_{\mathrm{m}}\end{pmatrix} \tag{28.28}$$

and

$$\begin{pmatrix}\rho_{\mathrm{e}}\\ \rho_{\mathrm{m}}\end{pmatrix} = \begin{bmatrix}\cos\xi & \sin\xi\\ -\sin\xi & \cos\xi\end{bmatrix}\cdot\begin{pmatrix}\rho_{\mathrm{e}}'\\ \rho_{\mathrm{m}}'\end{pmatrix} \tag{28.29}$$

It can be easily shown that under this duality transformation, Maxwell's equations are covariant. In other words, if the primed fields satisfy Maxwell's equations with the primed sources, the nonprimed fields satisfy Maxwell's equations with the nonprimed sources. Moreover, critical quantities quadratic in the fields, such as $\mathbf{E} \times \mathbf{H}$ and $\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H}$ remain invariant under this duality transformation (for real ξ), for example, $\mathbf{E} \times \mathbf{H} = \mathbf{E}' \times \mathbf{H}'$ [6]. As a special case of this property, for $\xi = -\pi/2$, we have the following correspondence:

$$\mathbf{E} \leftrightarrow -\mathbf{H}$$

$$\mathbf{H} \leftrightarrow \mathbf{E}$$

$$\mathbf{J}_\mathrm{m} \leftrightarrow \mathbf{J}_\mathrm{e}$$

$$\rho_\mathrm{m} \leftrightarrow \rho_\mathrm{e}$$

$$\mu \leftrightarrow \varepsilon$$

$$\varepsilon \leftrightarrow \mu$$

It is worth mentioning here that since the fields themselves cannot be measured and since all the important quantities (such as the power and energy terms, $\mathbf{E} \times \mathbf{H}$ and $\mathbf{E} \cdot \mathbf{D} + \mathbf{B} \cdot \mathbf{H}$) remain invariant under duality, it is just a matter of convenience which of several possible dual solutions to an electromagnetic problem we consider. This is precisely the reason why in some antenna problems, for example, one can formulate them with electric currents on the conductors, or with magnetic currents on the apertures and obtain the same detectable (measurable) results.

28.6.3 Lorentz Reciprocity Theorem

Consider the fields $\mathbf{E}_1$ and $\mathbf{H}_1$, generated by the system of sources $\mathbf{J}_{\mathrm{e}1}$ and $\mathbf{J}_{\mathrm{m}1}$ and the fields $\mathbf{E}_2$ and $\mathbf{H}_2$, generated by the system of sources $\mathbf{J}_{\mathrm{e}2}$ and $\mathbf{J}_{\mathrm{m}2}$. Then it can be shown that [8]

$$\oint_\mathrm{S} (\mathbf{E}_1 \times \mathbf{H}_2 - \mathbf{E}_2 \times \mathbf{H}_1) \cdot \mathrm{dS} = \int_\mathrm{V} (\mathbf{H}_1 \cdot \mathbf{J}_{\mathrm{m}2} - \mathbf{E}_1 \cdot \mathbf{J}_{\mathrm{e}2} - \mathbf{H}_2 \cdot \mathbf{J}_{\mathrm{m}1} + \mathbf{E}_2 \cdot \mathbf{J}_{\mathrm{e}1}) \mathrm{dV} \tag{28.30}$$

As a special case of this theorem, consider the volume of interest V and two sources in it, $\mathbf{J}_{\mathrm{e}1}$ and $\mathbf{J}_{\mathrm{e}2}$. When $\mathbf{J}_{\mathrm{e}1}$ is present alone, it generates the field $\mathbf{E}_1$. When $\mathbf{J}_{\mathrm{e}2}$ is present alone, it generates the field $\mathbf{E}_2$. Lorentz' reciprocity theorem as stated by Equation 28.30, implies that

$$\mathbf{E}_1 \cdot \mathbf{J}_{\mathrm{e}2} = \mathbf{E}_2 \cdot \mathbf{J}_{\mathrm{e}1} \tag{28.31}$$

The reciprocity theorem is used frequently in electromagnetic problem studies to facilitate solution of complicated problems. One well-known application of reciprocity has been in the expressions of the mutual and self-impedance of simple radiators. Reciprocity also manifests itself in some linear circuits so that Equation 28.13 holds when the electric fields and the current densities are replaced by voltages and currents, respectively. For details of this property see the text by Valkenburg [13].

28.6.4 Equivalent Principles (Theory of Images)

Equivalent principles are used to describe the electromagnetic field problem in a volume of interest in more than one configurations of sources. While the volume of interest is the same in all the equivalent configurations (materials, geometry, and source distributions), the geometry outside the volume of interest is different for each equivalent configuration. A detailed discussion of Field Equivalence Principles can be found in Collin [8]. Here only the method of images is mentioned. The method of images is based on the uniqueness of the field. Namely, as long as the sources inside the volume of interest, V, and the boundary conditions on $S_1, S_2, \ldots$ remain the same, the fields inside V are unique regardless of what sources exist outside V.

The method of images is based on this principle. Its applicability is limited to just a few canonical geometries. Consider, for example, a current carrying wire over an infinite perfect electric conductor,

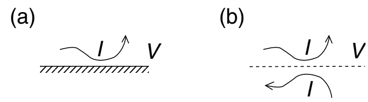


FIGURE 28.3 The fields in V are identical for both configurations (a) and (b). The fields outside V vanish in (a) while they are certainly nonzero in (b).

as shown in Figure 28.3a. The tangential electric field is zero on the perfect electric conductor (PEC) ground plane. Next consider the configuration shown in Figure 28.3b where the ground plane has been removed and the volume below it has a second current carrying wire, shaped in the mirror image of the original and with an opposite direction for the current. It can be shown with simple symmetry arguments that the tangential electric field on the dashed line (where the ground plane used to be) in configuration Figure 28.3b vanishes. Therefore, since Figure 28.3a and b are identical within V and they have the same boundary conditions (zero tangential electric field) on the boundary, the fields inside V are the same.

Clearly, it is much easier to numerically evaluate the solution in configuration (Figure 28.3b) than that of configuration (Figure 28.3a). Images are very often used in electromagnetic problems when appropriate. It is often easy to extend the theory of images to current carrying wires in the presence of corners of ground planes. Moreover, there have been several studies to extend the theory of images to more complex geometries [11]. However, the complexity of the images quickly escalates diminishing the benefits of removing the ground planes.

28.7 Simple Solution to Maxwell's Equations I (Unbounded Plane Waves)

In electromagnetics, a homogeneous and isotropic region of space with no sources in it is called free space. In free space, $\overline{\overline{\mu}} = \mu_o$, $\overline{\overline{\varepsilon}} = \varepsilon_o$ and Equations 28.6a through (d), can be decoupled by increasing the order of differentiation giving rise to the following second order wave equations:

$$\nabla^2 \mathbf{E} + \omega^2 \mu_o \varepsilon_o \mathbf{E} = 0 \tag{28.32}$$

$$\nabla^2 \mathbf{H} + \omega^2 \mu_o \varepsilon_o \mathbf{H} = 0 \tag{28.33}$$

Applying separation of variables, Equations 28.32 and 28.33 are found to have plane wave solutions of the form

$$\mathbf{E} = \mathbf{E}_o^+ \mathrm{e}^{-\mathrm{j}\mathbf{k}\cdot\mathbf{r}} + \mathbf{E}_o^- \mathrm{e}^{+\mathrm{j}\mathbf{k}\cdot\mathbf{r}} \tag{28.34}$$

$$\mathbf{H} = \mathbf{H}_o^+ \mathrm{e}^{-\mathrm{j}\mathbf{k}\cdot\mathbf{r}} + \mathbf{H}_o^- \mathrm{e}^{+\mathrm{j}\mathbf{k}\cdot\mathbf{r}} \tag{28.35}$$

The electromagnetic wave propagating in the $+\mathbf{k}$ direction ($\mathrm{e}^{-\mathrm{jk.r}}$) travels with the speed of light, $1/(\mu_o \varepsilon_o)$ and has a wave impedance,

$$\left|\frac{\mathbf{E}_o^+}{\mathbf{H}_o^+}\right| = \sqrt{\frac{\mu_o}{\varepsilon_o}}. \tag{28.36}$$

The wavenumber, k, obeys the relation

$$k = \frac{\omega}{\mathrm{c}} = \omega\sqrt{\mu_o \varepsilon_o} \tag{28.37}$$

where c is the speed of light. It can be shown that the power density carried by the plane wave is along the direction of propagation, **k**, and its average value over one period is given by the Poynting vector

$$\mathbf{S}^{+} = \frac{1}{2}\mathbf{E}_{\mathrm{o}}^{+} \times \mathbf{H}_{\mathrm{o}}^{+} \tag{28.38}$$

where the fields **E** and **H** are represented by their amplitude values in the phasor convention. It is worth noting that the Poynting vector itself is not a phasor. Being the product of two harmonic signals, the power density represented by the Poynting vector has a steady (DC) component and a harmonic component of twice the frequency of the electromagnetic field. The DC component of the Poynting vector is exactly given by Equation 28.38. Similar discussions hold for the wave propagating in the $-\mathbf{k}$ direction. These wave solutions obey all the usual wave phenomena of reflection and refraction when incident on interfaces between two different media. In general, the harmonic term of the Poynting vector is related to the reactive power flow in space. This power is required in order to satisfy the boundary conditions of the problem and it represents a continuous exchange between energy stored in the electric and the magnetic fields. For detailed discussions on these phenomena and more general cases of wave solutions to Maxwell's equations, the reader is referred to the following sections of this chapter and the text by Collin [8].

28.8 Simple Solution to Maxwell's Equations II (Guided Plane Waves)

It is possible to have electromagnetic waves inside cavities and waveguides. A waveguide is any structure which supports waves traveling along one direction and confined in the transverse plane by its boundaries. A rectangular duct made out of a conductor and a coaxial cable are waveguides with closed boundaries. A twin lead transmission line, or a dielectric plated conductor are among the many open boundary waveguides. Waves supported by waveguides are special solutions to Maxwell's equations.

Let us consider a waveguide along the z-axis. We are looking for solutions to Maxwell's equations that represent waves traveling along z. In other words, we are looking for solutions in the form:

$$\mathbf{E} = \mathbf{E}(\mathrm{x},\mathrm{y})\mathrm{e}^{-\mathrm{j}\beta \mathrm{z}} = \left(\mathbf{E}_{\mathrm{t}}(\mathrm{x},\mathrm{y}) + \hat{\mathbf{z}}\mathrm{E}_{\mathrm{z}}(\mathrm{x},\mathrm{y})\right)\mathrm{e}^{-\mathrm{j}\beta \mathrm{z}} \tag{28.39}$$

and

$$\mathbf{H} = \mathbf{H}(\mathrm{x},\mathrm{y})\mathrm{e}^{-\mathrm{j}\beta \mathrm{z}} = \left(\mathbf{H}_{\mathrm{t}}(\mathrm{x},\mathrm{y}) + \hat{\mathbf{z}}\mathrm{H}_{\mathrm{z}}(\mathrm{x},\mathrm{y})\right)\mathrm{e}^{-\mathrm{j}\beta \mathrm{z}} \tag{28.40}$$

where $\mathbf{E}_{\mathrm{t}}$ and $\mathbf{H}_{\mathrm{t}}$ are vectors in the transverse, (x, y) plane. Solving Equation 28.6, in a region of space free from any sources and looking specifically for solutions in the form of Equations 28.39 and 28.40 reduces the problem to the following wave equations:

$$\nabla_{\mathrm{t}}^{2}\mathrm{E}_{\mathrm{z}} + (\omega^{2}\mu\varepsilon - \beta^{2})\mathrm{E}_{\mathrm{z}} = 0 \tag{28.41}$$

and

$$\nabla_{\mathrm{t}}^{2}\mathrm{H}_{\mathrm{z}} + (\omega^{2}\mu\varepsilon - \beta^{2})\mathrm{H}_{\mathrm{z}} = 0 \tag{28.42}$$

The transverse fields are given in terms of the axial components of the waves as

$$\mathbf{E}_{\mathrm{t}}(\mathrm{x},\mathrm{y}) = \frac{1}{\mathrm{k}^{2} - \beta^{2}}(-\mathrm{j}\beta\nabla_{\mathrm{t}}\mathrm{E}_{\mathrm{z}}(\mathrm{x},\mathrm{y}) + \mathrm{j}\omega\mu\hat{\mathbf{z}} \times \nabla\mathrm{H}_{\mathrm{z}}(\mathrm{x},\mathrm{y})) \tag{28.43}$$

and

$$\mathbf{H}_t(x, y) = \frac{1}{k^2 - \beta^2}(-j\beta\nabla_t H_z(x, y) - j\omega\varepsilon\hat{\mathbf{z}} \times \nabla E_z(x, y)) \tag{28.44}$$

∇_t in the above equations stands for the transverse del operator, that is,

$$\nabla_t = \hat{\mathbf{x}}\frac{\partial}{\partial x} + \hat{\mathbf{y}}\frac{\partial}{\partial y}.$$

The wavenumber of the guided wave, β, is, in general, different from the wavenumber in the medium ($k^2 = \omega^2\mu\varepsilon$). The wavenumber is determined when the boundary condition specific to the waveguide are applied. There are some special cases of these guided wave solutions that are examined at a later section in this chapter in much more detail for practical microwave waveguides.

TEM modes (E and H transverse to the direction of propagation)

$$E_z = H_z = 0; \quad \beta^2 = k^2 = \omega^2\mu\varepsilon \quad \text{and} \quad \mathbf{E}_t = \sqrt{\frac{\mu}{\varepsilon}}(\mathbf{H}_t \times \hat{\mathbf{z}})$$

TM modes (H transverse to the direction of propagation)

$$E_z \neq 0; \quad H_z = 0$$

$$\nabla_t^2 E_z + (\omega^2\mu\varepsilon - \beta^2)E_z = 0 \quad \mathbf{E}_t(x, y) = \frac{-j\beta}{k^2 - \beta^2}\nabla_t E_z(x, y) \quad \text{and} \quad \mathbf{H}_t = -\frac{\omega\varepsilon}{\beta}\mathbf{E}_t \times \hat{\mathbf{z}}$$

TE modes (E transverse to the direction of propagation)

$$H_z \neq 0; \quad E_z = 0$$

$$\nabla_t^2 H_z + (\omega^2\mu\varepsilon - \beta^2)H_z = 0 \quad \mathbf{H}_t(x, y) = \frac{-j\beta}{k^2 - \beta^2}\nabla_t H_z(x, y) \quad \text{and} \quad \mathbf{E}_t = \frac{\omega\mu}{\beta}\mathbf{H}_t \times \hat{\mathbf{z}}$$

References

1. J.A. Stratton and L.J. Chu, *Phys. Rev.*, **56**, 99–107 (1939).
2. Samuel Silver, Ed., *Microwave Antenna Theory and Design*, chapter 3, McGraw Hill, 1949.
3. G. Arfken, *Mathematical Methods for Physicists*, 2nd edition, pp. 66–70, Academic Press, 1970.
4. W. Pauli, *Theory of Relativity*, Dover, NY, 1981.
5. J.D. Jackson, *Classical Electrodynamics*, 2nd edition, pp. 226–235, John Wiley & Sons, 1975.
6. J.D. Jackson, *Classical Electrodynamics*, 2nd edition, pp. 251–260, John Wiley & Sons, 1975.
7. C. Balanis, *Advanced Engineering Electromagnetics*, John Wiley & Sons, 1989.
8. R.E. Collin, *Field Theory of Guided Waves*, 2nd edition, IEEE Press, 1991.
9. J.A. Stratton, *Electromagnetic Theory*, McGraw Hill, 1941.
10. R.F. Harrington, *Time-Harmonic Electromagnetic Fields*, McGraw Hill, 1961.
11. I.V. Lindel, Image Theory for the Soft and Hard Surface, *IEEE Trans. Antenn. Propagat.*, **43**, 1 (January 1995).
12. R.P. Feynman, R.B. Leighton and M. Sands, "Electromagnetic Mass," in *The Feynman Lectures on Physics*, chapter 28, Addison Wesley, 1964.
13. M.E. Valkenburg, *Network Analysis*, 3rd edition, pp. 255–259, Prentice Hall, 1974.

29

Wave Propagation in Free Space

Matthew Sadiku
Prairie View A&M University

Sudarshan Rao Nelatury
Penn State University, Erie

The concept of propagation refers to the various ways by which an electromagnetic (EM) wave travels from the transmitting antenna to the receiving antenna. Propagation of EM wave may also be regarded as a means of transferring energy or information from one point (a transmitter) to another (a receiver). The transmission of analog or digital information from one point to another is the largest application of microwave frequencies. Therefore, understanding the principles of wave propagation is of practical interest to microwave engineers. Engineers cannot completely apply formulas or models for microwave system design without an adequate knowledge of the propagation issue.

Wave propagation at microwave frequencies has a number of advantages (Veley, 1987). First, microwaves can accommodate very wide bandwidth without causing interference problems because microwave frequencies are so high. Consequently, a huge amount of information can be handled by a single microwave carrier. Second, microwaves propagate along a straight line like light rays and are not bent by the ionosphere as are lower frequency signals. This straight-line propagation makes communication satellites possible. In essence, a communication satellite is a microwave relay station that is used in linking two or more grounded-based transmitters and receivers. Third, it is feasible to design highly directive antenna systems of a reasonable size at microwave frequencies. Fourth, compared with low-frequency electromagnetic waves, microwave energy is more easily controlled, concentrated, and directed. This makes it useful for cooking, drying, and physical diathermy. Moreover, the microwave spectrum provides more communication channels than the radio and TV bands. With the ever-increasing demand for channel allocation, microwave communication has become more common.

EM wave propagation is achieved through guided structures such as transmission lines and waveguides or through space. In this chapter, our major focus is on EM wave propagation in free space and the power resident in the wave.

EM wave propagation can be described by two complimentary models. The physicist attempts a theoretical model based on universal laws, which extends the field of application more widely than currently known. The engineer prefers an empirical model based on measurements, which can be used immediately. This chapter presents the complimentary standpoints by discussing theoretical factors affecting wave propagation and the semiempirical rules allowing handy engineering calculations. First, we consider wave propagation in idealistic simple media, with no obstacles. We later consider the more realistic case of wave propagation around the earth, as influenced by its curvature and by atmospheric conditions.

29.1 Wave Equation

The conventional propagation models, on which the basic calculation of microwave links is based, result directly from Maxwell's equations (Sadiku, 2007):

$$\nabla \cdot \mathbf{D} = \rho_v \tag{29.1}$$

$$\nabla \cdot \mathbf{B} = 0 \tag{29.2}$$

$$\nabla \times \mathbf{E} = -\frac{\partial \mathbf{B}}{\partial t} \tag{29.3}$$

$$\nabla \times \mathbf{H} = \mathbf{J} + \frac{\partial \mathbf{D}}{\partial t} \tag{29.4}$$

In these equations, **E** is electric field strength in volts per meter, **H** is magnetic field strength in amperes per meter, **D** is electric flux density in coulombs per square meter, **B** is magnetic flux density in webers per square meter, **J** is conduction current density in amperes per square meter, and ρ_v is electric charge density in coulombs per cubic meter. These equations go hand in hand with the constitutive equations for the medium:

$$\mathbf{D} = \epsilon \mathbf{E} \tag{29.5}$$

$$\mathbf{B} = \mu \mathbf{H} \tag{29.6}$$

$$\mathbf{J} = \sigma \mathbf{E} \tag{29.7}$$

where $\epsilon = \epsilon_o \epsilon_r$, $\mu = \mu_o \mu_r$, and σ are the permittivity, the permeability, and the conductivity of the medium, respectively.

Consider the general case of a lossy medium that is charge-free ($\rho_v = 0$). Assuming time-harmonic fields and suppressing the time factor $e^{j\omega t}$, Equations 29.1 through 29.7 can be manipulated to yield Helmholtz's wave equations

$$\nabla^2 \mathbf{E} - \gamma^2 \mathbf{E} = 0 \tag{29.8}$$

$$\nabla^2 \mathbf{H} - \gamma^2 \mathbf{H} = 0 \tag{29.9}$$

where $\gamma = \alpha + j\beta$ is the *propagation constant,* α is the *attenuation constant* in nepers per meter and β is the *phase constant* in radians per meter. Constants α and β are given by

$$\alpha = \omega \sqrt{\frac{\mu\epsilon}{2}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2} - 1\right]} \tag{29.10}$$

$$\beta = \omega \sqrt{\frac{\mu\epsilon}{2}\left[\sqrt{1+\left(\frac{\sigma}{\omega\epsilon}\right)^2} + 1\right]} \tag{29.11}$$

where $\omega = 2\pi f$ is the frequency of the wave. The wavelength λ and wave velocity u are given in terms of β as

$$\lambda = \frac{2\pi}{\beta} \tag{29.12}$$

$$u = \frac{\omega}{\beta} = f\lambda \tag{29.13}$$

Without loss of generality, if we assume that wave propagates in the z-direction and the wave is polarized in the x-direction, solving the wave equations 29.8 and 29.9 results in

$$\mathbf{E} = E_o e^{-\alpha z}\cos(\omega t - \beta z)\mathbf{a}_x \tag{29.14}$$

where $\mathbf{a}_x$ is the unit vector in the x-direction.

$$\mathbf{H} = \frac{E_o}{|\eta|} e^{-\alpha z}\cos(\omega t - \beta z - \theta_\eta)\mathbf{a}_y \tag{29.15}$$

where $\mathbf{a}_y$ is the unit vector in the y-direction and $\eta = |\eta|\angle\theta_\eta$ is the *intrinsic impedance* of the medium and is given by

$$|\eta| = \frac{\sqrt{\frac{\mu}{\epsilon}}}{\sqrt[4]{\left[1+\left(\frac{\sigma}{\omega\epsilon}\right)^2\right]}}, \quad \tan 2\theta_\eta = \frac{\sigma}{\omega\epsilon}, \quad 0 \le \theta_\eta \le 45^\circ \tag{29.16}$$

Equations 29.14 and 29.15 show that as the EM wave propagates in the medium, its amplitude is attenuated according to $e^{-\alpha z}$, as illustrated in Figure 29.1. The distance δ through which the wave amplitude is reduced by a factor of e^{-1} (about 37%) is called the *skin depth* or *penetration depth* of the medium, that is,

$$\delta = \frac{1}{\alpha} \tag{29.17}$$

The power density of the EM wave is obtained from the Poynting vector

$$\mathbf{P} = \mathbf{E} \times \mathbf{H} \tag{29.18}$$

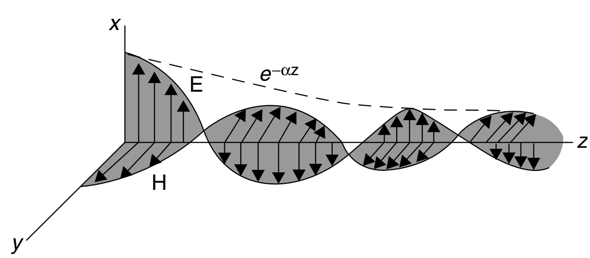


FIGURE 29.1 The magnetic and electric field components of a plane wave in a lossy medium.

with the time-average value of

$$\begin{aligned} P_{\text{ave}} &= \text{Re}\left(\mathbf{E} \times \mathbf{H}^*\right) \\ &= \frac{E_o^2}{2|\eta|} e^{-2\alpha z} \cos\theta_\eta \mathbf{a}_z \end{aligned} \tag{29.19}$$

It should be noted from Equations 29.14 and 29.15 that **E** and **H** are everywhere perpendicular to each other and also to the direction of wave propagation. Thus, the wave described by Equations 29.14 and 29.15 is said to be *plane polarized,* implying that the electric field is always parallel to the same plane (the *xz* plane in this case) and is perpendicular to the direction of propagation. Also, as mentioned earlier, the wave decays as it travels in the *z* direction because of loss. This loss is expressed in the *complex relative permittivity* of the medium.

$$\epsilon_c = \epsilon_r' - j\epsilon_r'' = \epsilon_r\left(1 - j\frac{\sigma}{\omega\epsilon}\right) \tag{29.20}$$

and measured by the *loss tangent,* defined by

$$\tan\delta = \frac{\epsilon_r''}{\epsilon_r'} = \frac{\sigma}{\omega\epsilon} \tag{29.21}$$

The imaginary part $\epsilon_r'' = \sigma/\omega\epsilon_o$ corresponds to the losses in the medium. The refractive index of the medium n is given by

$$n = \sqrt{\epsilon_r} \tag{29.22}$$

Having considered the general case of wave propagation through a lossy medium, we now consider wave propagation in other types of media. A medium is said to be a good conductor if the loss tangent is large ($\sigma \gg \omega\epsilon$) or a lossless or good dielectric if the loss tangent is very small ($\sigma \ll \omega\epsilon$). Thus, the characteristics of wave propagation through other types of media can be obtained as special cases of wave propagation in a lossy medium as follows:

1. Good conductors: $\sigma \gg \omega\epsilon$, $\epsilon = \epsilon_o$, $\mu = \mu_o\mu_r$
2. Good dielectrics: $\sigma \ll \omega\epsilon$, $\epsilon = \epsilon_o$, $\mu = \mu_o\mu_r$
3. Free space: $\sigma = 0$, $\epsilon = \epsilon_o$, $\mu = \mu_o$

where $\epsilon_o = 8.854 \times 10^{-12}$ F/m is the free-space permittivity, and $\mu_o = 4\pi \times 10^{-7}$ H/m is the free-space permeability. The conditions for each medium type are merely substituted in Equations 29.10 to 29.21 to obtain the wave properties for that medium.

The classical model of wave propagation presented in this section helps us understand some basic concepts of EM wave propagation and the various parameters that play a part in determining the progress of the wave from the transmitter to the receiver. We will apply the ideas to the particular case of wave propagation in free space or the atmosphere in the section on Propagation in the Atmosphere. Before then, we digress a little and consider the important issue of wave polarization.

29.2 Wave Polarization

The concept of polarization is an important property of an EM wave that has been developed to described the various types of electric field variation and orientation. It is therefore a common practice to describe an EM wave by its polarization. The polarization of an EM wave depends on the transmitting antenna or source. It is determined by the direction of the electric field. It is regarded as the locus of the tip of the electric field (in a plane perpendicular to the direction of propagation) at a given point in space as a function of time. For this reason, there are four types of polarization: linear or plane, circular, elliptic, and random.

In linear or plane polarized waves, the orientation of the field is constant in space and time. For a plane traveling in the $+z$ direction, the electric field may be written as

$$\mathbf{E}(z,t) = E_x(z,t)\mathbf{a}_x + E_y(z,t)\mathbf{a}_y \tag{29.23}$$

where

$$E_x = \mathrm{Re}\left[E_{ox}e^{j(\omega t - kz + \phi_x)}\right] = E_{ox}\cos(\omega t - kz + \phi_x) \tag{29.24a}$$

$$E_y = \mathrm{Re}\left[E_{oy}e^{j(\omega t - kz + \phi_x)}\right] = E_{oy}\cos(\omega t - kz + \phi_y) \tag{29.24b}$$

For linear polarization, the phase difference between the x and y components must be

$$\Delta\phi = \phi_y - \phi_x = n\pi, \quad n = 0, 1, 2, \ldots \tag{29.25}$$

This allows the two components to maintain the same ratio at all times, which implies that the electric field always lies along a straight line in a constant $-z$ plane. In other words, if we observe the wave in the direction of propagation (z in this case), we will notice that the tip of the electric field follows a line. Hence, the term *linear polarization*. Linearly polarized plane waves can be generated by simple antennas (such as dipole antennas) or lasers.

Circular polarized waves are characterized by an electric field with constant magnitude and orientation rotating in a plane transverse to the direction of propagation. Circular polarization takes place when the x and y components are the same in magnitude ($E_x = E_y$) and the phase difference between them is an odd multiple of $\pi/2$, that is,

$$\Delta\phi = \phi_y - \phi_x = \pm\,(2n+1)\frac{\pi}{2}, \quad n = 0, 1, 2, \ldots \tag{29.26}$$

The resultant vector field rotates around the axis of propagation as a function of time and space. Circularly polarized waves can be generated by a helically wound wire antenna or by two linear sources that are oriented perpendicular to each other and fed with currents that are out of phase by 90°.

Linear and circular polarizations are special cases of the more general case of the elliptical polarization. An elliptically polarized wave is one in which the tip of the field traces an elliptic locus in a fixed transverse plane as the field changes with time. Elliptical polarization is achieved when appropriately defined x and y components are not equal in magnitude ($E_x \neq E_y$) and the phase difference between them is an odd multiple of $\pi/2$, i.e.,

$$\Delta\phi = \phi_y - \phi_x = \pm\,(2n+1)\frac{\pi}{2}, \quad n = 0, 1, 2, \ldots \tag{29.27}$$

This allows the tip of the electric field to trace an ellipse in the x–y plane.

The polarized waves described so far are illustrated in Figure 29.2. They are deterministic meaning that the field is a predictable function of space and time. If the field is completely random, the wave is said to be randomly polarized. Typical examples of such waves are radiation from the sun and radio stars.

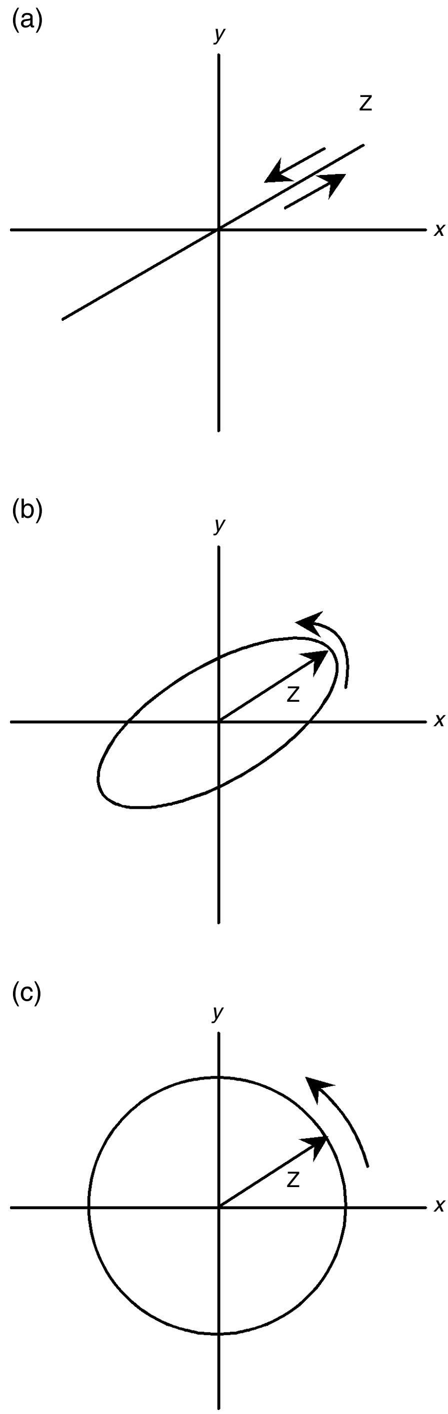


FIGURE 29.2 Wave polarizations: (a) linear, (b) elliptic, (c) circular.

29.3 Propagation in the Atmosphere

Wave propagation hardly occurs under the idealized conditions assumed previously. For most communication links, the previous analysis must be modified to account for the presence of the earth, the ionosphere, and atmospheric precipitates such as fog, raindrops, snow, and hail. This is done in this section.

The major regions of the earth's atmosphere that are of importance in radio wave propagation are the troposphere and the ionosphere. At radar frequencies (approximately 100 MHz to 300 GHz), the troposphere is by far the most important. It is the lower atmosphere comprised of a nonionized region extending from the earth's surface up to about 15 km. The ionosphere is the earth's upper atmosphere in the altitude region from 50 km to one earth radius (6370 km). Sufficient ionization exists in this region to influence wave propagation.

Wave propagation over the surface of the earth may assume any of the following three principal modes:

- Surface wave propagation along the surface of the earth;
- Space wave propagation through the lower atmosphere;
- Sky wave propagation by reflection from the upper atmosphere.

These modes are portrayed in Figure 29.3. The sky wave is directed toward the ionosphere, which bends the propagation path back toward the earth under certain conditions in a limited frequency range (few to 50 MHz approximately). This is highly dependent on the condition of the ionosphere (its level of ionization) and the signal frequency. The surface (or ground) wave takes effect at the low-frequency end of the spectrum (2 to 5 MHz approximately) and is directed along the surface over which the wave is propagated. Since the propagation of the ground wave depends on the conductivity of the earth's surface, the wave is attenuated more than if it were propagation through free space. For frequencies above 50 MHz, space wave propagation is the best possible mode. The space wave consists of the direct wave and the reflected wave. The direct wave travels from the transmitter to the receiver in nearly a straight path while the reflected wave is due to ground reflection. The space wave obeys the optical laws in that direct and reflected wave components contribute to the total wave component. Although the sky and surface waves are important in many applications, we will only consider the space wave in this chapter.

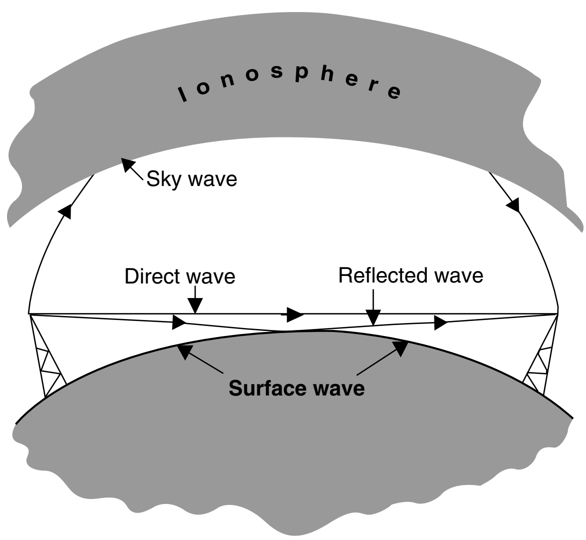


FIGURE 29.3 Modes of wave propagation.

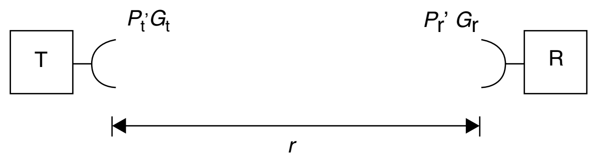


FIGURE 29.4 Transmitting and receiving antennas in free space.

Figure 29.4 depicts the electromagnetic energy transmission between two antennas in space. As a wave radiates from the transmitting antenna and propagates in space, its power density decreases, as expressed ideally in Equation 29.19. Assuming that the antennas are in a lossless medium or free space, the power received by the receiving antenna is given by the *Friis transmission equation* (Liu and Fang, 1988):

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t \tag{29.28}$$

where the subscripts t and r respectively refer to transmitting and receiving antennas. In Equation 29.28, P = power in watts, G = antenna gain (dimensionless), r = distance between the antennas in meters, and λ = wavelength in meters. The Friis equation relates the power received by one antenna to the power transmitted by the other, provided that the two antennas are separated by $r > 2D^2/\lambda$, and D is the largest dimension of either antenna. Thus, the Friis equation applies only when the two antennas are in the far field of each other. It shows that the received power decays at a rate of 20 dB/decade with distance. In case the propagation path is not in free space, a correction factor F is included to account for the effect of the medium. This factor, known as the *propagation factor,* is simply the ratio of the electric field intensity E_m in the medium to the electric field intensity E_o in free space, i.e.,

$$F = \frac{E_m}{E_o} \tag{29.29}$$

The magnitude of F is always less than unity since E_m is always less than E_o. Thus, for a lossy medium, Equation 29.28 becomes

$$P_r = G_r G_t \left(\frac{\lambda}{4\pi r}\right)^2 P_t |F|^2 \tag{29.30}$$

For practical reasons, Equations 29.28 or 29.29 are commonly expressed in logarithmic form. If all the terms are expressed in decibels (dB), Equation 29.30 can be written in logarithmic form as

$$P_r = P_t + G_r + G_t - L_o - L_m \tag{29.31}$$

where P = power in dB referred to 1 W (or simply dBW), G = gain in dB, L_o = free-space loss in dB, and L_m = loss in dB due to the medium. The free-space loss is obtained from standard nomograph or directly from

$$L_o = 20 \log\left(\frac{4\pi r}{\lambda}\right) \tag{29.32}$$

while the loss due to the medium is given by

$$L_m = -20 \log|F \qquad (29.33)$$

Our major concern in the rest of this section is to determine L_o and L_m for two important cases of space propagation that differ considerably from the free-space conditions.

29.4 Effect of the Earth

The phenomenon of multipath propagation causes significant departures from free-space conditions. The term *multipath* denotes the possibility of an EM wave propagating along various paths from the transmitter to the receiver. In multipath propagation of an EM wave over the earth's surface, two such paths exist: a direct path and a path via reflection and diffractions from the interface between the atmosphere and the earth. A simplified geometry of the multipath situation is shown in Figure 29.5.

In the preceding discussion it was noted that the magnitude of F in a homogenous lossy medium is always less than unity. However, if we consider the reflections from the Earth's surface, we need to redefine F. Without considering the losses in the medium and assuming that the reflection coefficient is $\Gamma = |\Gamma|e^{j\phi}$, we can express the correction factor F as

$$F = 1 + \Gamma e^{-j\Delta} \qquad 29.34$$

where Δ is the phase difference corresponding to the path difference between direct and ground reflected ray. Assuming a flat and perfectly conducting Earth, for which $\Gamma = -1$, we can depict the variation in F in the form of a coverage diagram, which is a plot of relative field strength as a function of direction in space from the transmitting antenna. It is usually drawn for a fixed transmitting antenna height h_1 in free-space wavelengths λ_o, and a fixed free-space range r_f in the h_2-d plane, where h_2 is the receiver antenna height and d is the distance between the transmitter and the receiver. A typical coverage diagram looks as shown in Figure 29.6 for the choice of values r_f = 2 km, h_1 = 120 λ_o.

The wave incident on the ground is not only reflected but is also diffracted. The reflected and diffracted component is commonly separated into two parts: one *specular* (or coherent) and the other *diffuse* (or incoherent), that can be separately analyzed. The specular component is well defined in terms of its amplitude, phase, and incident direction. Its main characteristic is its conformance to Snell's law for reflection, which requires that the angles of incidence and reflection be equal and coplanar. It is a plane

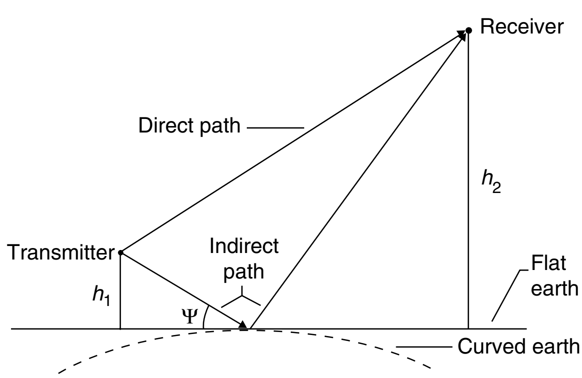


FIGURE 29.5 Multipath geometry.

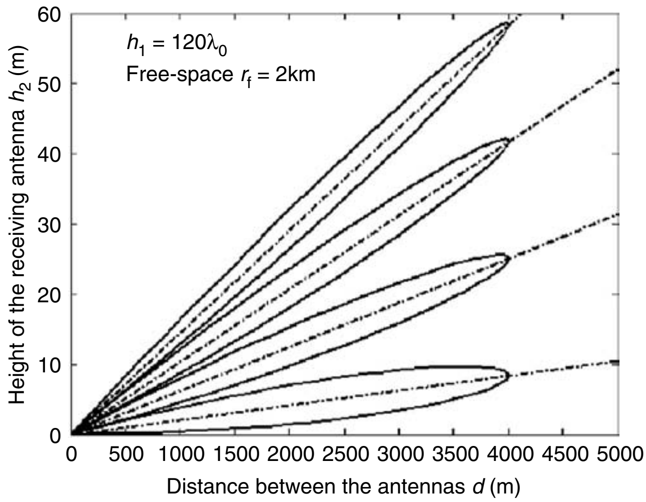


FIGURE 29.6 A typical coverage diagram for a flat perfectly conducting Earth.

wave, and as such, is uniquely specified by its direction. The diffuse component, however, arises out of the random nature of the scattering surface, and as such, is nondeterministic. It is not a plane wave and does not obey Snell's law for reflection. It does not come from a given direction but from a continuum.

The loss factor F that accounts for the departures from free-space conditions is given by

$$F = 1 + \Gamma \rho_s D S(\theta) e^{-j\Delta} \tag{29.35}$$

where Γ = Fresnel reflection coefficient,
ρ_s = roughness coefficient,
D = divergence factor,
$S(\theta)$ = shadowing function,
Δ = is the phase angle corresponding to the path difference

The Fresnel reflection coefficient Γ accounts for the electrical properties of the earth's surface. Since the earth is a lossy medium, the value of the reflection coefficient depends on the complex relative permittivity ϵ_c of the surface, the grazing angle ψ, and the wave polarization. It is given by

$$\Gamma = \frac{\sin\psi - z}{\sin\psi + z} \tag{29.36}$$

where

$$z = \sqrt{\epsilon_c - \cos^2\psi} \quad \text{for horizontal polarization,} \tag{29.37}$$

$$z = \frac{\sqrt{\epsilon_c - \cos^2\psi}}{\epsilon_c} \quad \text{for vertical polarization,} \tag{29.38}$$

$$\epsilon_c = \epsilon_r - j\frac{\sigma}{\omega\epsilon_o} = \epsilon_r - j60\sigma\lambda \tag{29.39}$$

As can be seen the reflection coefficient varies with polarization and the medium parameters. In Figure 29.7a–d we show plots of magnitude and phase of reflection coefficient in case of sea water at four typical values of wavelengths λ = 3 cm, 50 cm, 1 m, and 3 m. The figures show that both the magnitude and phase of the reflection coefficient vary significantly in case of vertical polarization rather than horizontal polarization.

To account for the spreading (or divergence) of the reflected rays due to earth curvature, we introduce the divergence factor *D*. The curvature has a tendency to spread out the reflected energy more than a corresponding flat surface. The divergence factor is defined as the ratio of the reflected field from a curved surface to the reflected field from a flat surface (Kerr, 1951). Using the geometry of Figure 29.8, *D* is given by

$$D \simeq \left(1 + \frac{2G_1G_2}{a_e G \sin\psi}\right)^{-1/2} \tag{29.40}$$

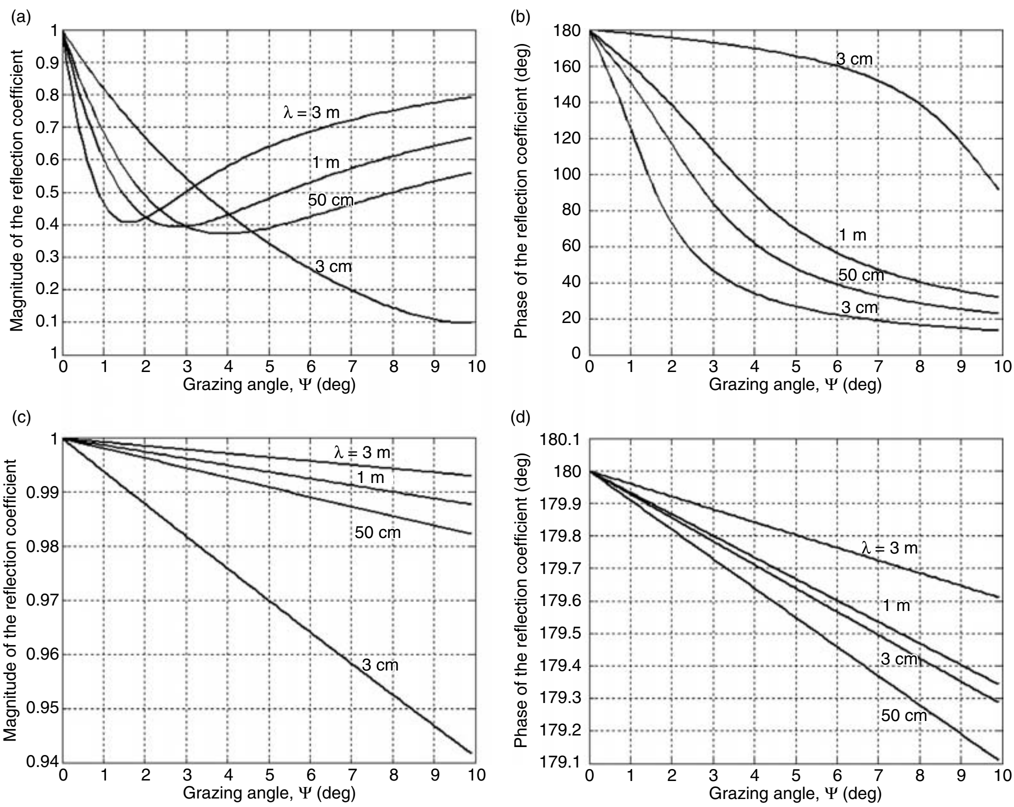


FIGURE 29.7 Variation in reflection coefficient in case of sea water surface. Values used for ε_c = 30-j420λ, for four representative values of λ = 50cm, 3cm, 1m and 3m. (a–b) Magnitude and phase of the reflection coefficient respectively in case of vertical polarization. (c–d) Magnitude and phase of the reflection coefficient respectively in case of horizontal polarization.

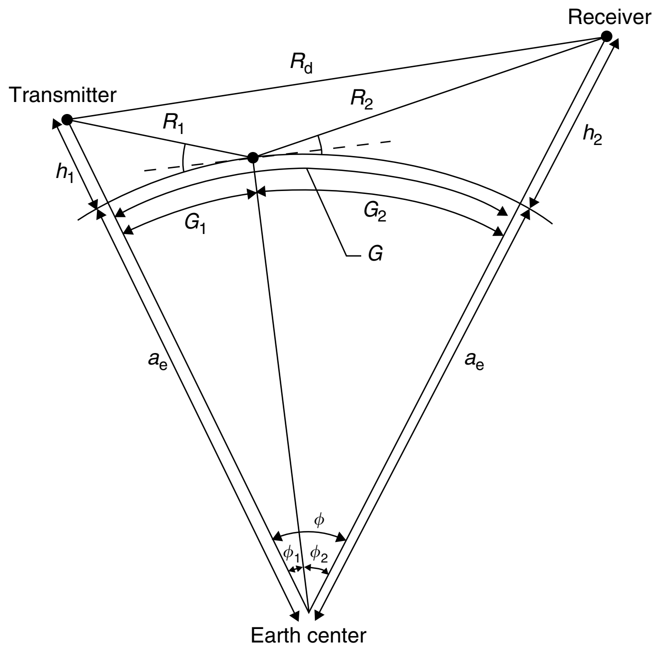


FIGURE 29.8 Geometry of spherical earth reflection.

where $G = G_1 + G_2$ is the total ground range and $a_e = 6370$ km is the effective earth radius. Given the transmitter height h_1, the receiver height h_2, and the total ground range G, we can determine G_1, G_2, and ψ. If we define

$$p = \frac{2}{\sqrt{3}}\left[a_e\left(h_1 + h_2\right) + \frac{G^2}{4}\right]^{1/2} \tag{29.41}$$

$$\alpha = \cos^{-1}\left[\frac{2a_e\left(h_1 - h_2\right)G}{p^3}\right] \tag{29.42}$$

and assume $h_1 \le h_2$, $G_1 \le G_2$, using small angle approximation yields

$$G_1 = \frac{G}{2} + p\cos\left(\frac{\pi + \alpha}{3}\right) \tag{29.43}$$

$$G_2 = G - G_1 \tag{29.44}$$

$$\phi_i = \frac{G_i}{a_e}, \quad i = 1, 2 \tag{29.45}$$

$$R_i = \left[h_i^2 + 4a_e\left(a_e + h_i\right)\sin^2\left(\phi_i/2\right)\right]^{1/2} \quad i = 1, 2 \tag{29.46}$$

The grazing angle is given by

$$\psi = \sin^{-1}\left[\frac{2a_e h_1 + h_1^2 - R_1^2}{2a_e R_1}\right] \tag{29.47}$$

or

$$\psi = \sin^{-1}\left[\frac{2a_e h_1 + h_1^2 + R_1^2}{2\left(a_e + h_1\right)R_1}\right] - \phi_1 \tag{29.48}$$

Although D varies from 0 to 1, in practice D is a significant factor at low grazing angle ψ(less than 0.1%).

The phase angle corresponding to the path difference between direct and reflected waves is given by

$$\Delta = \frac{2\pi}{\lambda}\left(R_1 + R_2 - R_d\right) \tag{29.49}$$

It would be interesting to redefine and obtain the coverage diagrams including the effects of Earth's curvature and diffraction from the ground but that requires some mathematical treatment and plotting on an orthogonal grid. Interested readers might consult (Collin, 1985).

The roughness coefficient ρ_s takes care of the fact that the earth's surface is not sufficiently smooth to produce specular (mirror-like) reflection except at a very low grazing angle. The earth's surface has a height distribution that is random in nature. The randomness arises out of the hills, structures, vegetation, and ocean waves. It is found that the distribution of the different heights on the earth's surface is usually the Gaussian or normal distribution of probability theory. If σ_h is the standard deviation of the normal distribution of heights, we define the roughness parameters as

$$g = \frac{\sigma_h \sin\psi}{\lambda} \tag{29.50}$$

If $g < 1/8$, specular reflection is dominant; if $g > 1/8$, diffuse scattering results. This criterion, known as the *Rayleigh criterion,* should only be used as a guideline since the dividing line between a specular and a diffuse reflection or between a smooth and a rough surface is not well defined. The roughness is taken into account by the roughness coefficient ($0 < \rho_s < 1$), which is the ratio of the field strength after reflection with roughness taken into account to that which would be received if the surface were smooth. The roughness coefficient is given by

$$\rho_s = \exp\left[-2\left(2\pi g\right)^2\right] \tag{29.51}$$

The shadowing function $S(\theta)$ is important at a low grazing angle. It considers the effect of geometric shadowing — the fact that the incident wave cannot illuminate parts of the earth's surface shadowed by higher parts. In a geometric approach, where diffraction and multiple scattering effects are neglected, the reflecting surface will consist of well-defined zones of illumination and shadow. As there will be no field on a shadowed portion of the surface, the analysis should include only the illuminated portions of the surface. The phenomenon of shadowing of a stationary surface was first investigated by Beckman in 1965 and subsequently refined by Smith (1967) and others. A pictorial representation of rough surfaces

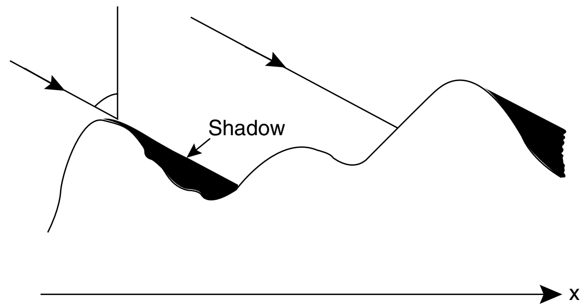


FIGURE 29.9 Rough surface illuminated at an angle of incidence θ.

illuminated at the angle of incidence θ (= 90° – ψ) is shown in Figure 29.9. It is evident from the figure that the shadowing function $S(\theta)$ equals unity when $\theta = 0$ and zero when $\theta = \pi/2$. According to Smith (1967),

$$S(\theta) \simeq \frac{\left[1 - \frac{1}{2}\,\mathrm{erfc}(a)\right]}{1 + 2B} \tag{29.52}$$

where erfc(x) is the complementary error function,

$$\mathrm{erfc}(x) = 1 - \mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2}\,dt \tag{29.53}$$

and

$$B = \frac{1}{4a}\left[\frac{1}{\sqrt{\pi}} e^{a^2} - a\,\mathrm{erfc}(a)\right], \tag{29.54}$$

$$a = \frac{\cot\theta}{2s}, \tag{29.55}$$

$$s = \frac{\sigma_h}{\sigma_l} = \text{rms surface slope} \tag{29.56}$$

In Equation 29.56, σ_h is the rms roughness height and σ_l is the correlation length. Alternative models for $S(\theta)$ are available in the literature. Using Equations 29.36 to 29.56, the loss factor in Equation 29.32 can be calculated. Thus

$$L_o = 20\,\log\left(\frac{4\pi R_d}{\lambda}\right) \tag{29.57}$$

$$L_m = -20\,\log\left(1 + \Gamma\,\rho_s\,D\,S(\theta)\,e^{-j\Delta}\right) \tag{29.58}$$

29.5 Effect of Atmospheric Hydrometeors

The effect of atmospheric hydrometeors on satellite-earth propagation is of major concern at microwave frequencies. The problem of scattering of electromagnetic waves by atmospheric hydrometeors has attracted much interest since the late 1940s. The main hydrometeors that exist for long duration, and have the greatest interaction with microwaves are rain, snow, and dust particles. At frequencies above 10 GHz, rain has been recognized as the most fundamental obstacle in the earth-space path. Rain has been known to cause attenuation, phase difference, and depolarization of radio waves. For analog signals, the effect of rain is more significant above 10 GHz while for digital signals, rain effects can be significant down to 3 GHz. Attenuation of microwaves due to precipitation becomes severe owing to increased scattering and beam energy absorption by raindrops thus impairing terrestrial as well as earth-satellite communication links. Cross-polarization distortion due to rain has also been engaging the attention of researchers. This is particularly of interest when frequency reuse employing signals with orthogonal polarizations are used for doubling the capacity of a communication system. Thorough reviews on the interaction of microwaves with hydrometeors have been done by Oguchi (1983).

The loss due to rain-filled medium is given by

$$L_m = \gamma(R)\,\ell_e(R)\,p(R) \tag{29.59}$$

where γ = attenuation per unit length at rain rate R,
ℓ_e = equivalent path length at rain rate R, and
$p(R)$ = probability in percentage of rainfall rate R.

The attenuation is a function of the cumulative rain-rate distribution, drop-size distribution, refractive index of water, temperature, and other variables. A rigorous calculation of $\gamma(R)$ using various numerical modeling tools and incorporating raindrop size distribution, velocity of raindrops, and the refractive index of water can be found in Sadiku (2001). For practical engineering purposes, what is needed is a simple formula relating attenuation to rain parameters. Such is found in the aR^b empirical relationship, which has been employed to calculate rain attenuation directly (Collin, 1985), that is,

$$\gamma(R) = aR^b \quad \text{dB/km} \tag{29.60}$$

where R is the rain rate and a and b are constants. At 0°C, the values of a and b are related to frequency f in gigahertz as follows:

$$a = G_a f^{E_a} \tag{29.61}$$

where

$$G_a = 6.39\times10^{-5}, \quad E_a = 2.03, \quad \text{for } f < 2.9 \text{ GHz}$$

$$G_a = 4.21\times10^{-5}, \quad E_a = 2.42, \quad \text{for } 2.9 \text{ GHz} \le f \le 54 \text{ GHz}$$

$$G_a = 4.09\times10^{-2}, \quad E_a = 0.699, \quad \text{for } 54 \text{ GHz} \le f < 100 \text{ GHz}$$

$$G_a = 3.38, \quad E_a = -0.151, \quad \text{for } 180 \text{ GHz} < f$$

and

$$b = G_b f^{E_b} \tag{29.62}$$

where

$$\begin{aligned}
&G_b = 0.851, \quad E_b = 0.158, &&\text{for } f < 8.5 \text{ GHz}\\
&G_b = 1.41, \quad E_b = -0.0779, &&\text{for } 8.5 \text{ GHz} \le f < 25 \text{ GHz}\\
&G_b = 2.63, \quad E_b = -0.272, &&\text{for } 25 \text{ GHz} \le f < 164 \text{ GHz}\\
&G_b = 0.616, \quad E_b = 0.0126, &&\text{for } 164 \text{ GHz} \le f.
\end{aligned}$$

The effective length $\ell_e(R)$ through the medium is needed since rain intensity is not uniform over the path. Its actual value depends on the particular area of interest and therefore has a number of representations (Liu and Fang, 1988). Based on data collected in western Europe and eastern North America, the effective path length has been approximated as (Hyde, 1984)

$$\ell_e(R) = \left[0.00741R^{0.766} + (0.232 - 0.00018R)\sin\theta\right]^{-1} \tag{29.63}$$

where θ is the elevation angle.

The cumulative probability in percentage of rainfall rate R is given by (Hyde, 1984)

$$p(R) = \frac{M}{87.66}\left[0.03\beta e^{-0.03R} + 0.2(1-\beta)\left(e^{-0.258R} + 1.86e^{-1.63R}\right)\right] \tag{29.64}$$

where M is mean the annual rainfall accumulation in mm and β is the Rice–Holmberg thunderstorm ratio.

The effect of other hydrometeors such as vapor, fog, hail, snow, and ice is governed by fundamental principles similar to the effect of rain (Collin, 1985). However, their effects are at least an order of magnitude less than the effect of rain in most cases.

29.5.1 Recent Developments

Existing literature on radiowave propagation and smart antennas for wireless communications carries information on empirical, deterministic and numerical models to predict the path loss in outdoor environments due to irregular terrain and presence of buildings (Janaswamy, 2000). Some of the empirical models are based on curve-fitting measurement results. Without taking into account the actual terrain topography or building database in the prediction step, they provide quick and reasonable statistical answers. Site-specific or deterministic models, on the other hand, do take into account the details of local terrain, but are slow in computation. Some popular examples in this category are ray-based methods (Bertoni, 2000), method of moments (MOM) for VHF propagation (Johnson et al., 1997) and fast finite difference methods (Janaswamy, 1994) etc. More recently, some formulations have considered the parabolic approximation to the Helmholtz equation, which assume forward scattering, wherein fields are coupled in the principal direction but not in the opposite direction (Levy 2000). An example is the knife-edge diffraction. The parabolic equation yields very useful results for long-distance propagation problems. The seminal idea of the parabolic equation dates back to the pioneering work of Leontovich and Fock (1965), which was popularized later by Tappert (1977). A common technique used for modeling EM wave propagation in the troposphere by a parabolic-type equation is the Fourier split-step algorithm. This algorithm permits one to solve a variety of complex propagation problems including the three-dimensional rural and urban terrains. A complete 3-dimensional vector formulation of the parabolic equation is found in (Janaswamy, 2003). This method is known to reduce the mean error relative to the actual measurements, to 1dB and standard deviation to less than 5 dB.

Recently, Sevgi et al. (2005) have developed a MATLAB-based package called SSPE_GUI, which can be used directly in simulations of short and long-range radiowave propagation over non-smooth Earth surfaces above which, exists an inhomogeneous atmosphere. The package is a graphical user interface (GUI) and is developed based on a two-dimensional split-step parabolic-equation (SSPE); thus the name SSPE_GUI. This package is simpler and easy to use especially for educational purposes. Moreover, the user may easily build arbitrary terrain profiles. The package may be downloaded from http://www3.dogus.edu.tr/lsevgi. We shall give some salient points of this software below.

The standard parabolic wave equation derived from the two-dimensional Helmholtz equation is given by

$$\frac{\partial^2 u}{\partial x^2} + 2jk_0 \frac{\partial u}{\partial z} + k_0^2\left(n^2 - 1\right)u = 0 \tag{29.65}$$

where $u(x, z)$ denotes the wave amplitude, and x and z stand for transverse and longitudinal coordinates, respectively. When the direction of propagation is predominantly along the z axis, one might separate the rapidly varying phase term to obtain an amplitude factor, which varies slowly with respect to the range z. In Equation (29.65) n is the refractive index, and k_0 is the free-space wavenumber. Applying the Fourier transform from the x domain to the spectral k_x domain, the above equation can be rewritten as a first-order ordinary differential equation as follows. Let the Fourier transform of $u(x, z)$ be

$\mathcal{F}\{u(z,x)\} = U(z,k_x)$. Noting that the operator $\frac{\partial^2}{\partial x^2}$ would be modified as $-k_x^2$ under Fourier transformation, Equation 29.65 becomes

$$\frac{\mathrm{d}U(z,k_x)}{\mathrm{d}z} + \frac{1}{2jk_0}\left[k_0^2(n^2-1) - k_x^2\right]U(z,k_x) = 0 \tag{29.66}$$

The solution of the above is straightforward and is set down as

$$u(z,x) = \exp\left[j\frac{k_0}{2}\left(n^2-1\right)\Delta z\right]\mathcal{F}^{-1}\left\{\exp\left[-\frac{jk_x^2\Delta z}{2k_0}\right]\mathcal{F}\left\{u(z_0,x)\right\}\right\} \tag{29.67}$$

Here the refractive index of the medium, n, is treated as a constant. Although in real problems n can be a function of height and/or range i.e., $n = n(z,x)$, this approach is sill acceptable, because the equation is solved at each small range step size, $\Delta\ z$, which is chosen small enough so that within any Δz interval, the refractive index is essentially constant. One can then calculate $u(z, x)$ along z in steps of Δz, starting with the knowledge of the initial field distribution, $u(z_0, x)$. The functional details of SSPE_GUI and what the user should know to start doing the simulations are fully covered in (Sevgi et al. 2005).

The following are the user defined inputs required for obtaining a visual picture of the wave propagation are: (1) Operating frequency in MHz, (2) Maximum desired range in km (3) Height of the transmitter (4) Observation Range step size (5) Atmospheric refractivity described by the first height up

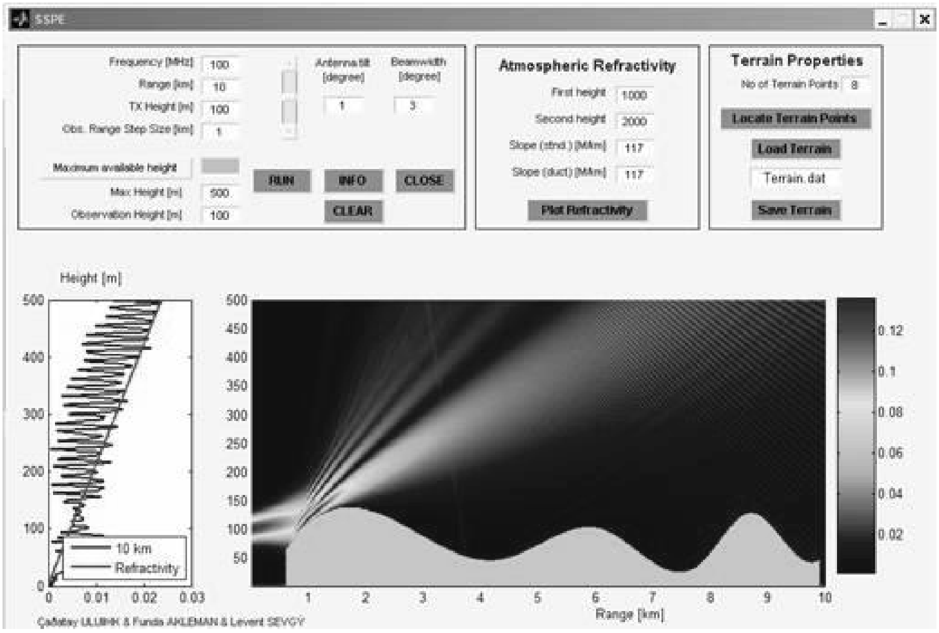


FIGURE 29.10 Showing a smaple screen-capture obtained by running the SSPE_GUI for the user-fed parameters that are appearing in the frames.

to which the atmosphere is standard, and above which lies an elevated duct (6) Second Height bordering the duct (7) Slope 1 (M/km) which is the slope of the standard atmosphere from ground to first height (8) Slope 2 (M/km) being the slope of duct from first height and second height (9) Maximum desired height (10) Antenna tilt in degrees (11) Beamwidth in degrees for antenna pattern (12) Number of terrain points that are going to be located for the terrain profile by the user. (13) Terrain geometry described by either a data file. Alternatively, there is even a provision for the user to input a chosen number of landmark points of the terrain by left-clicking the mouse button in the graphic window; whereupon, the program builds the terrain by spline interpolation. For convenience, we provide in Figure 29.10 a sample screen capture for the parameters that appear in the frames.

The SSPE methods or the implicit finite difference methods are microscopic in nature, in that, the differential equation is applied to a microscopic region. Some researchers also proposed macroscopic propagators that give results rather fast. However, they address a restricted class of boundary conditions and for a general scenario still require sound justification. One should consider more advanced mesh structures and absorbing boundary conditions and devise ways to take into account back scattering also especially if a receiver is located behind tall structures.

29.6 Other Effects

Besides hydrometeors, the atmosphere has the composition given in Table 29.1. While attenuation of EM waves by hydrometeors may result from both absorption and scattering, gases act only as absorbers. Although some of these gases do not absorb microwaves, some possess permanent electric and/or magnetic dipole moment and play some part in microwave absorption. For example, nitrogen molecules do not possess permanent electric or magnetic dipole moment and therefore play no part in microwave absorption. Oxygen has a small magnetic moment that enables it to display weak absorption lines in the centimeter- and millimeter-wave regions. Water vapor is a molecular gas with a permanent electric dipole moment. It is more responsive to excitation by an EM field than is oxygen.

TABLE 29.1 Composition of Dry Atmosphere from Sea Level to about 90 km (Livingston, 1970)

Constituent	Percent by Volume	Percent by Weight
Nitrogen	78.088	75.527
Oxygen	20.949	23.143
Argon	0.93	1.282
Carbon dioxide	0.03	0.0456
Neon	1.8×10^{-3}	1.25×10^{-3}
Helium	5.24×10^{-4}	7.24×10^{-5}
Methane	1.4×10^{-4}	7.75×10^{-5}
Krypton	1.14×10^{-4}	3.30×10^{-4}
Nitrous oxide	5×10^{-5}	7.6×10^{-5}
Xenon	8.6×10^{-6}	3.90×10^{-5}
Hydrogen	5×10^{-5}	3.48×10^{-6}

Other mechanisms that can affect EM wave propagation in free space, not discussed in this chapter, include clouds, dust, and the ionosphere. The effect of the ionosphere is discussed in detail in standard texts.

References

Collin, R. E., 1985. *Antennas and Radiowave Propagation,* McGraw-Hill, New York, 339–456.

Tappert, F. D. "The Parabolic Approximation Method," in J. B. Keller and J. S. Papadakis (eds.) Wave Propagation and Under-

Freeman, R. L., 1994. *Reference Manual for Telecommunications Engineering,* 2nd ed., John Wiley & Sons, New York, 711–768.

Bertoni, H. L. Radio Propagation for ModernWireless Systems. Upper Saddle River, NJ: Prentice Hall PTR, 2000.

Johnson et al., J. T. "A method of moments model for VHF propagation," IEEE Trans Antennas Propagat., vol. 45, pp. 115–125, Jan. 1997.

Hyde, G., 1984. Microwave propagation, in R. C. Johnson and H. Jasik (eds.), *Antenna Engineering Handbook,* 2nd ed., McGraw-Hill, New York, 45.1–45.17.

Kerr, D. E., 1951. *Propagation of Short Radio Waves,* McGraw-Hill, New York, (Republished by Peter Peregrinus, London, U.K., 1987), 396–444.

Liu, C. H. and D. J. Fang, 1988. Propagation, in Y. T. Lo and S. W. Lee, *Antenna Handbook: Theory, Applications, and Design,* Van Nostrand Reinhold, New York, 29.1–29.56.

Livingston, D. C., 1970. *The Physics of Microwave Propagation,* Prentice-Hall, Englewood Cliffs, NJ, 11.

Sevgi, L., C.Uluisik, and F.Akleman, "A MATLAB-based two-dimensional parabolic equation rasdiowave propagation package," IEEE Antenna and Propagation Magazine, vol. 47, no. 4, pp. 164–175, Aug. 2005,

Levy, M. F. *Parabolic Equation Methods for Electromagnetic WavePropagation.* London, U.K.: IEE Press, 2000.

Oguchi, T., 1983. Electromagnetic Wave Propagation and Scattering in Rain and Other Hydrometeors, *Proc. IEEE,* 71, 1029–1078.

Janaswamy, R. "A fast finite difference method for propagation predictions irregular, inhomogeneous terrain," IEEE Trans. Antennas Propagat., vol. 42, pp. 1257–1267, Sept. 1994.

Janaswamy, R. "Path Loss Predictions in the Presence of Buildings on Flat Terrain: A 3-D Vector Parabolic Equation approach," IEEE Trans. Antennas Propagat., vol. 51, pp. 1716–1728, Aug., 2003.

Janaswamy, R. Radiowave Propagation & Smart Antennas for Wireless Communications. Boston, MA: Kluwer, 2000.

Sadiku, M. N. O., 2007. *Elements of Electromagnetics,* 4th ed. Oxford University Press, New York, Chapter 10, 429–499.

Sadiku, M. N. O., 2001. *Numerical Techniques in Electromagnetics,* 2nd ed., CRC Press, Boca Raton, FL, 95–105.
Smith, B. G., 1967. Geometrical shadowing of a random rough surface, *IEEE Trans. Ant. Prog.,* 15, 668–671.
Fock, V. A., *Electromagnetic Diffraction and Propagation Problems*, Oxford, Pergamon, 1965.
Veley, V. F., *Modern Microwave Technology,* Prentice-Hall, Englewood Cliffs, NJ, 1987, 23–33.

Further Information

The subject of wave propagation could easily fill many chapters, and here it has only been possible to point out some of the main points of concern to microwave systems engineer. Today 60 GHz radio is gaining much attention. There are several sources of information dealing with the theory and practice of wave propagation in space. Some of these are in the references section. Journals such as *Radio Science, IEE Proceedings Part H, IEEE Transactions on Antenna and Propagation* are devoted to EM wave propagation. *Radio Science* is available at American Geophysical Union, 2000 Florida Avenue, NW, Washington DC 20009; *IEE Proceedings Part H* at IEE Publishing Department, Michael Faraday House, 6 Hills Way, Stevenage, Herts SG1 2AY, U.K.; and *IEEE Transactions on Antenna and Propagation* at IEEE, 445 Hoes Lane, P. O. Box 1331, Piscataway, NJ 08855-1331.

30

Guided Wave Propagation and Transmission Lines

W.R. Deal
V. Radisic
Northrop Grumman Corporation

Y. Qian
Microsemi Corporation

T. Itoh
University of California

At higher frequencies where wavelength becomes small with respect to feature size, it is often necessary to consider an electronic signal as an electromagnetic wave and the structure where this signal exists as a waveguide. A variety of different concepts can be used to examine this wave behavior. The most simplistic view is transmission line theory, where propagation is considered in a one-dimensional (1-D) manner and the cross-sectional variation of the guided wave is entirely represented in terms of distributed transmission parameters in an equivalent circuit. This is the starting point for transmission line theory that is commonly used to design microwave circuits. In other guided wave structures, such as enclosed waveguides, it is more appropriate to examine the concepts of wave propagation from the perspective of Maxwell's equations, the solutions of which will explicitly demonstrate the cross-sectional dependence of the guided wave structure.

Most practical wave guiding structures rely on single-mode propagation, which is restricted to a single direction. This allows the propagating wave to be categorized according to its polarization properties. A convenient method is classifying the modes as transverse electromagnetic (TEM), transverse electric (TE), or transverse magnetic (TM). TEM modes have both the electric and magnetic field transverse to the direction of propagation. Only the magnetic field is transverse to the direction of propagation in TM modes, and only the electric field is transverse to the direction of propagation in TE modes.

In this chapter, we first briefly examine the telegrapher's equations, which are the starting point for transmission line theory. The simple transmission line model accurately describes a number of guided wave structures and is the starting point for transmission line theory. In the next section, enclosed waveguides including rectangular and circular waveguide will be discussed. Relevant concepts such as cut-off frequency and modes will be given. In the final section, four common planar guided wave structures will be discussed. These inexpensive and compact structures are the foundation for the modern commercial RF front-end.

30.1 TEM Transmission Lines, Telegrapher's Equations, and Transmission Line Theory

In this section, the concept of guided waves in simple TEM-guiding structures will be explored in terms of the simple model provided by the Telegrapher's equations, also referred to as the transmission line equations. The Telegrapher's equations demonstrate the guided wave properties in terms of lumped equivalent circuit parameters available for many types of simple two-conductor transmission lines, and are valid for all types of TEM waveguide if their corresponding equivalent circuit parameters are known. These parameters must be found from Maxwell's equation in their fundamental form. Finally, properties and parameters for several types of two-wire TEM transmission line structures are introduced.

A transmission line or waveguide is used to transmit power and information from one point to another in an efficient manner. Three common types of transmission lines, which support TEM guided waves, are shown in Figure 30.1a through c, including parallel-plate transmission line, two-wire line, and coaxial transmission line. The first, parallel plate transmission line consists of a dielectric slab sandwiched between two parallel conducting plates of width, w. More practical, commonly used variations of this structure at microwave and millimeter-wave frequencies include microstrip and stripline, which will be briefly discussed in the final section of this chapter. Shown in Figure 30.1b is two-wire transmission line, consisting of two parallel conducting lines separated by a distance d. This is commonly used for power distribution at low frequencies. Finally, coaxial transmission line consists of two concentric conductors separated by a dielectric layer. This structure is well shielded and commonly used at high frequencies well into microwave frequencies.

The telegrapher's equations form a simple and intuitive starting point for the physics of guided wave propagation in these structures. Figure 30.2 shows an equivalent circuit model for a two-conductor transmission line of differential length Δz in terms of the following four parameters:

R—resistance per unit length of both conductors (Ω/m)
L—inductance per unit length of both conductors (H/m)
G—conductance per unit length (S/m)
C—capacitance per unit length of both conductors (F/m)

These parameters represent physical quantities for each of the relevant transmission lines. For each of the structures shown in Figure 30.1a through c, R represents conductor losses, L represents inductance, G represents dielectric losses, and C represents the capacitance between the two lines.

Returning to Figure 30.2, the quantities $v(z,t)$ and $v(z+\Delta z,t)$ represent change in voltage along the differential length of transmission line, while $i(z,t)$ and $i(z+\Delta z,t)$ represent the change in current. Writing Kirchoff's voltage law and current laws for the structure, dividing by Δz, and applying the fundamental theorem of calculus as $\Delta z \to 0$, two differential coupled differential equations known as the telegrapher's equations are obtained:

$$-\frac{\partial v(z,t)}{\partial z} = Ri(z,t) + L\frac{\partial i(z,t)}{\partial t} \tag{30.1}$$

$$-\frac{\partial i(z,t)}{\partial z} = Gi(z,t) + C\frac{\partial v(z,t)}{\partial t} \tag{30.2}$$

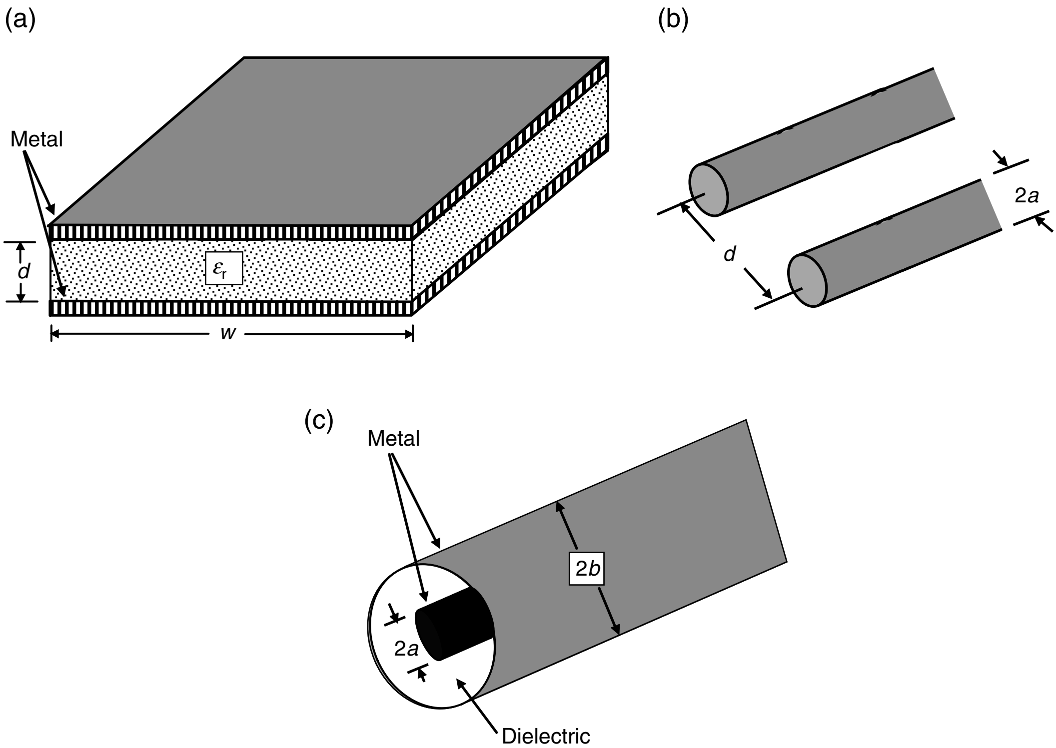


FIGURE 30.1 Three simple TEM-type transmission line geometries: (a) parallel-plate transmission line, (b) two-wire line, and (c) coaxial line.

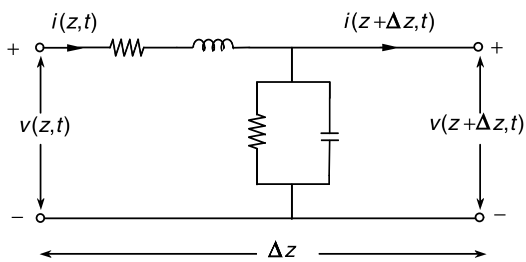


FIGURE 30.2 Distributed equivalent circuit model for a transmission line.

However, typically we are interested in signals with harmonic time-dependence ($e^{j\omega t}$). In this case, the time-harmonic forms of the telegrapher's equations are given by

$$-\frac{\mathrm{d}V(z)}{\mathrm{d}z} = (R + j\omega L)I(z) \tag{30.3}$$

$$-\frac{\mathrm{d}I(z)}{\mathrm{d}z} = (G + j\omega C)V(z) \tag{30.4}$$

The constant γ is defined to be the propagation constant with real and imaginary parts, α and β, corresponding to the attenuation constant (Np/m) and phase constant (rad/m) in the following manner:

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega C)} \tag{30.5}$$

This may then be substituted into the telegrapher's equations, which may then be solved for $V(z)$ and $I(z)$ to yield the following one-dimensional wave equations:

$$\frac{d^2V(z)}{dz^2} - \gamma^2 V(z) = 0 \tag{30.6}$$

$$\frac{d^2I(z)}{dz^2} - \gamma^2 I(z) = 0 \tag{30.7}$$

The form of this equation is the well-known wave equation. This indicates that the transmission line will support a guided electromagnetic wave traveling along the z-direction. The telegrapher's equations use a physical equivalent circuit and basic circuit theory to demonstrate the wave behavior of an electromagnetic signal on a transmission line. Alternatively, the same result can be obtained by starting directly with Maxwell's equations in their fundamental form, which may be used to derive the wave equation for a propagating electromagnetic wave. In this case, the solution of the wave equation will be governed by the boundary conditions. Similarly, the parameters R, L, G, and C are determined by the geometry of the transmission line structures.

Returning to the telegrapher's equations, several important facts may be noted. First, the characteristic impedance of the transmission line may be found by taking the ratio of the forward traveling voltage and current wave amplitudes and is given in terms of the equivalent circuit parameters as

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{30.8}$$

In the case of a lossless transmission line, this reduces to $Z_0 = \sqrt{L/C}$. The phase velocity, also known as the propagation velocity, is the velocity of the wave as it moves along the waveguide. It is defined to be

$$v_p = \frac{\omega}{\beta} \tag{30.9}$$

In the lossless case, this reduces to

$$v_p = \frac{1}{\sqrt{LC}} = \frac{1}{\sqrt{\mu\varepsilon}} \tag{30.10}$$

This shows that the velocity of the signal is directly related to the medium. In the case of an air-filled purely TEM mode, the wave will propagate at the familiar value $c = 3 \times 10^8$ m/s. In addition, it provides a relationship between L, C, and the medium that the wave is guided in. Therefore, if the properties of the medium are known, it is only necessary to determine to determine either L or C. Once C is known, G may be determined by the following relationship:

$$\frac{G}{C} = \frac{\sigma}{\varepsilon} \tag{30.11}$$

Note that σ is the conductivity of the medium and not of the metal conductors. The final parameter, the series resistance R, is determined by the power loss in the conductors. Simple approximations for the transmission line parameters R, L, G, and C for the three types of transmission lines shown in Figure 30.1a through c are well known and are shown in Table 30.1. Note that, μ, ε, and σ relate to the medium separating the conductors, and σ_c refers to the conductor. Once the equivalent circuit parameters are determined, the characteristic impedance and propagation constant of the transmission line may be

TABLE 30.1 Transmission Line Parameters for Parallel-Plate, Two-Wire Line, and Coaxial Transmission Lines

	Parallel-Plate Waveguide	Two-Wire Line	Coaxial Line
R (Ω/m)	$\frac{2}{w}R_s$	$\frac{R_s}{\pi a}$	$\frac{R_s}{2\pi}\left(\frac{1}{a}+\frac{1}{b}\right)$
L (H/m)	$\mu\frac{d}{w}$	$\frac{\mu}{\pi}\cosh^{-1}\left(\frac{D}{2a}\right)$	$\frac{\mu}{2\pi}\ln\left(\frac{b}{a}\right)$
G (S/m)	$\sigma\frac{w}{d}$	$\frac{\pi\sigma}{\cosh^{-1}(D/2a)}$	$\frac{2\pi\sigma}{\ln(b/a)}$
C (F/m)	$\varepsilon\frac{w}{d}$	$\frac{\pi\varepsilon}{\cosh^{-1}(D/2a)}$	$\frac{2\pi\varepsilon}{\ln(b/a)}$

determined. Note that R_s represents the surface resistance of the conductors, given as

$$R_s = \sqrt{\frac{\pi f \mu_c}{\sigma_c}} \tag{30.12}$$

In general, the loss characteristics of the conductor are characterized by its skin depth. The skin depth is the distance into the conductor necessary for the electric field amplitude to decay by a factor of 1/**e**, or 38.6%. To minimize conductor losses, the conductor of a transmission line must be several skin depths thick. A plot of electric field amplitude decay from the surface of the conductor is shown in Figure 30.3a. Note that the electric field has fallen by more than 99% in a thickness of five skin depths. Since the field amplitude decays to zero inside the conductor, additional conductor does not lower the RF conductor loss once the RF current has decayed to a sufficiently low level. The skin depth for gold in microns (μm) is illustrated in Figure 30.3b. At 1 GHz, the skin depth is 7.9 μm thick. To minimize RF conductor losses, a five-skin depth conductor would have to be 40-μm thick or almost 2 mils. At 100 GHz, the skin depth is only 0.25-μm thick and a good conductor thickness for low conductor losses would be 1.25-μm thick, which is compatible with standard MMIC metallization schemes. At RF frequencies, however, printed circuit board technologies typically offer the lowest passive circuit losses since they offer considerably thicker metallizations.

30.2 Guided Wave Solution from Maxwell's Equations, Rectangular Waveguide, and Circular Waveguide

A waveguide is any structure that guides an electromagnetic wave. In the preceding section, several simple TEM transmission structures were discussed. While these structures do support a guided wave, the term waveguide more commonly refers to a closed metallic structure with a fixed cross section within which a guided wave propagates, as shown for the arbitrary cross section in Figure 30.4. The guide is filled with a material of permittivity ε and permeability μ, and is defined by its metallic wall parallel to the z-axis. These structures demonstrate lower losses than the simple transmission line structures of the first section and are used to transport power in the microwave and millimeter wave frequency range. Ohmic losses are low and the waveguide is capable of carrying large power levels. Disadvantages are bulk, weight, and limited bandwidth, which cause planar transmission lines to be used wherever possible in modern communications circuits. However, a wide variety of components are available in this technology, including high-performance filters, couplers, isolators, attenuators, and detectors.

Inside this type of an enclosed waveguide, an infinite number of distinct solutions exist, each one of these solutions being referred to as a *waveguide mode*. At a given operating frequency, the cross section

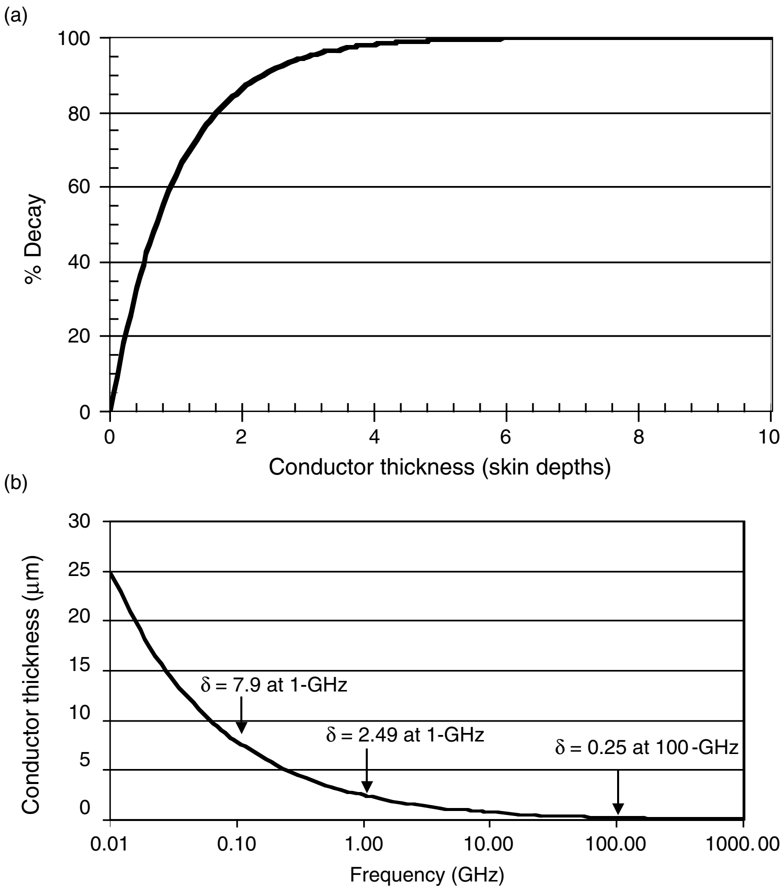


FIGURE 30.3 Illustration of skin depth effect including (a) field amplitude decay inside a conductor in terms of skin depths and (b) skin depth of gold as a function of frequency.

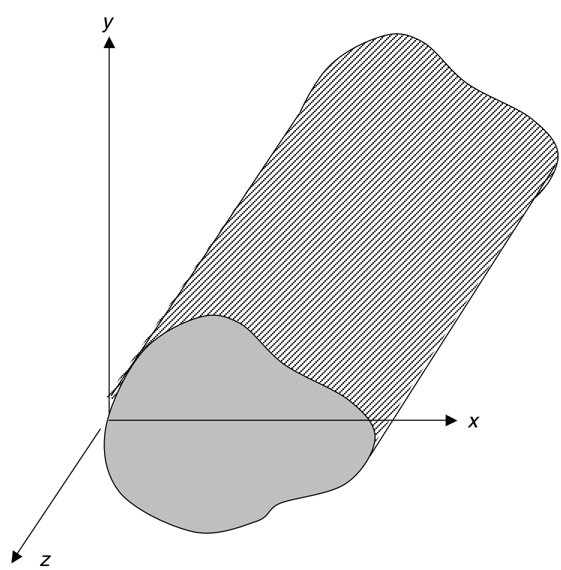


FIGURE 30.4 Geometry of enclosed waveguide with arbitrary cross section. Propagation is in the z-direction.

of the waveguide and the type of material in the waveguide determine the characteristics of these modes. These modes are usually classified by the longitudinal components of the electric and magnetic field, E_z and H_z, respectively, where propagation is in the z-direction. The most common classifications are TE, TM, EH, and HE modes. The basic characteristics will be described in the next two paragraphs. The TEM modes that were discussed in the previous section do not propagate in this type of metallic enclosed waveguide. This is because a TEM mode requires two conductors for it to propagate, where conventional enclosed waveguide has only a single enclosing conductor.

The two most common waveguide modes are the TE and TM modes. TE modes have no component of E in the z-direction, which means that E is completely transverse to the direction of propagation. Similarly, TM modes have no component of H in the z-direction.

EH and HE modes are hybrid modes which may be present under certain conditions, such as a waveguide partially filled with dielectric. In this case, pure TE and TM are unable to satisfy all of the necessary boundary conditions and a more complex type of modal solution is required. With both EH and HE, neither E nor H in the direction of propagation are zero. In EH modes, the characteristics of the transverse fields are controlled more by H_z rather than by E_z. HE modes are controlled more by E_z rather than by H_z. These types of hybrid modes may also be referred to as longitudinal section electric (LSE) and longitudinal section magnetic (LSM). It should be noted that most commonly used waveguides are homogenous; being entirely filled with material of a single permittivity (which may, of course, be air) and these types of modes will not be present.

Inside a homogenous waveguide, E_z and H_z satisfy the scalar wave equation inside the waveguide:

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) E_z + h^2 E_z = 0 \tag{30.13}$$

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) H_z + h^2 H_z = 0 \tag{30.14}$$

Note that h is given as

$$h^2 = \omega^2 \mu\varepsilon + \gamma^2 = k^2 + \gamma^2 \tag{30.15}$$

The wave number, k, is for the material filling the waveguide. For several simple and homogenous waveguides commonly used waveguide geometries, these equations may be solved to obtain closed form solutions by applying boundary equations on the walls of the waveguide. The resulting modal solution will possess distinct eigenvalues determined by the cross section of the waveguide. One important result obtained from this procedure is that waveguide modes, unlike the fundamental TEM mode that propagates in two-wire structures at any frequency, will have a distinct cut-off frequency. It may be shown that the propagation constant varies with frequency as

$$\gamma = \alpha + j\beta = h\sqrt{1 - \left(\frac{f}{f_c}\right)^2} \tag{30.16}$$

where the cut-off frequency, f_c, is given by:

$$f_c = \frac{h}{2\pi\sqrt{\mu\varepsilon}} \tag{30.17}$$

By inspection of Equation 30.16, and recalling the $\exp(j\omega t - \gamma z)$ dependence of the wave propagating in the $+z$ direction (for propagation in the $-z$ direction, replace z with $-z$), the physical significance of the cut-off frequency is clear. For a given mode, when $f > f_c$, the propagation constant γ is imaginary and the wave is propagating. Alternatively, when $f < f_c$, the propagation constant γ is real and the wave decays exponentially. In this case, modes operated below cut-off frequency attenuate rapidly and are therefore referred to as evanescent modes. In practice, a given waveguide geometry is seldom operated at a frequency where more than one mode will propagate. This fixes the bandwidth of the waveguide to operate at some point above the cut-off frequency of the fundamental mode and below the cut-off frequency of the second-order mode. Although in some rare instances, higher-order modes may be used for specialized applications.

The guided wavelength is also a function of the cross section geometry of the waveguide structure. The guided wavelength is given as

$$\lambda_g = \frac{\lambda_0}{\sqrt{1 - (f_c/f)^2}} \tag{30.18}$$

Note that λ_0 is the wavelength of a plane wave propagating in an infinite medium of the same material as the waveguide. Two important facts may be noted about this expression. First, at frequencies well above the cut-off frequency, $\lambda_g \approx \lambda$. Second, as $f \rightarrow f_c$, $\lambda \rightarrow \infty$, further illustrating that the mode does not propagate. This is another reason that the operating frequency is always chosen above the cut-off frequency. This concept is graphically depicted in Figure 30.5, a β/k diagram for standard WR-90 waveguide. At the cut-off frequency, the phase constant goes to zero, indicating that the wave does not propagate. At high frequencies, β approaches the phase constant in an infinite region of the same medium. Therefore, β/k approaches one.

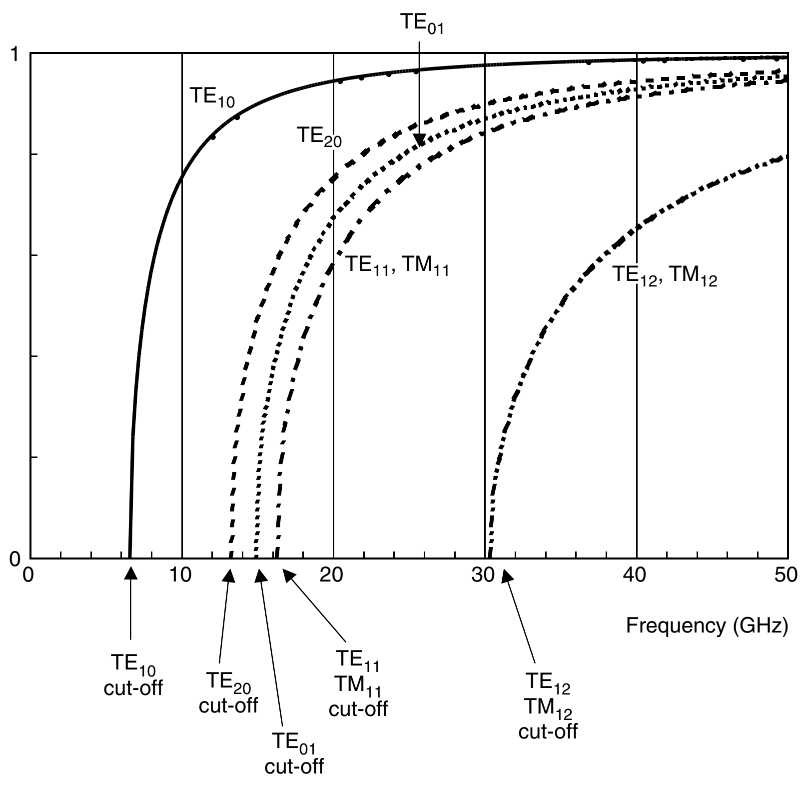


FIGURE 30.5 β/k diagram for WR-90 waveguide illustrating the concept of higher mode propagation and cut-off frequency.

The wave impedance of waveguide is given by the ratio of the magnitudes of the TE and TM field components, which will be constant across the cross section of the waveguide. For a given mode, the wave impedance for the TE and TM modes are given as

$$Z_{TE} = \frac{E_T}{H_T} = \frac{j\omega\mu}{\gamma} \tag{30.19}$$

$$Z_{TM} = \frac{E_T}{H_T} = \frac{\gamma}{j\omega\varepsilon} \tag{30.20}$$

E_T and H_T represent the TE and TM fields. Note that, at frequencies well above cut-off, the wave impedance for both the TE and TM modes approaches $\sqrt{\mu/\varepsilon}$, the characteristic impedance of a plane wave propagating in an infinite medium of the same material as the waveguide. Further, as $f \to f_c$, then $Z_{TE} \to \infty$ and $Z_{TM} \to 0$, again demonstrating the necessity of choosing an operating point well above cut-off.

A variety of geometries are used for waveguide, the most common being rectangular waveguide, which is used in the microwave and well into the millimeter-wave frequency regime. As shown in Figure 30.6, it is a rectangular metallic guide of width a, and height b. Rectangular waveguide propagate both TE and TM modes. For conciseness, the field components of the TE_{mn} and TM_{mn} modes are presented in Table 30.2. From the basic form of the equations, we see that effect of the rectangular cross section is a standing wave-dependence determined by the dimensions of the cross section, a and b. Further, h (and therefore the propagation constant, γ) is determined by a and b. The dimensions of the waveguide are chosen so that only a single mode propagates at the desired frequency, with all other modes cutoff. By convention, $a > b$ and a ratio of $a/b = 2.1$ is typical for commercial waveguide types.

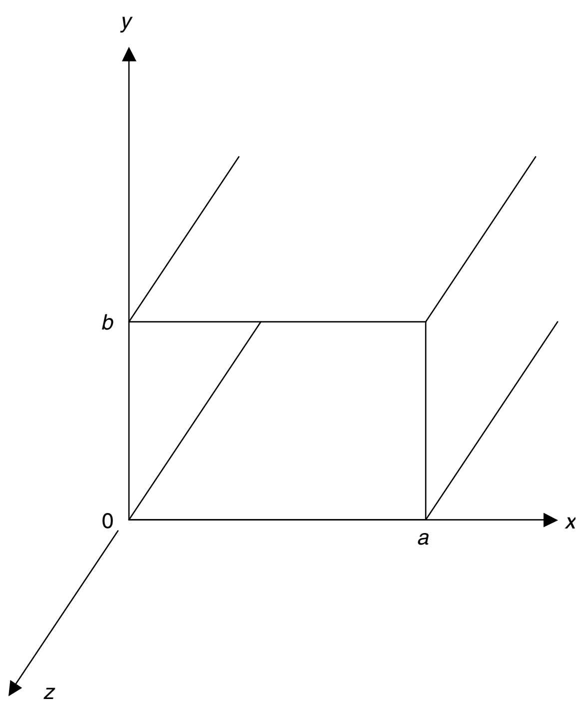


FIGURE 30.6 Geometry of rectangular waveguide.

TABLE 30.2 Field Components for Rectangular Waveguide

	TE	TM
E_z	0	$E_0 \sin\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$
H_z	$H_0 \cos\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$	0
E_x	$H_0 \frac{j\omega\mu n\pi}{h_{mn}^2 b} \cos\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$	$-E_0 \frac{\gamma_{mn} m\pi}{h_{mn}^2 a} \cos\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$
H_x	$H_0 \frac{\gamma_{mn} m\pi}{h_{mn}^2 a} \sin\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$	$H_0 \frac{j\omega\varepsilon n\pi}{h_{mn}^2 b} \sin\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$
E_y	$-H_0 \frac{j\omega\mu m\pi}{h_{mn}^2 a} \sin\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$	$-E_0 \frac{\gamma_{mn} n\pi}{h_{mn}^2 b} \sin\left(\frac{m\pi x}{a}\right)\cos\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$
H_y	$H_0 \frac{\gamma_{mn} n\pi}{h_{mn}^2 b} \cos\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$	$-E_0 \frac{j\omega\varepsilon m\pi}{h_{mn}^2 a} \cos\left(\frac{m\pi x}{a}\right)\sin\left(\frac{n\pi y}{b}\right) e^{-\gamma_{mn}z}$
h_{mn}	$\sqrt{\left(\frac{m\pi x}{a}\right)^2 + \left(\frac{n\pi y}{b}\right)^2} = 2\pi f_c\sqrt{\mu\varepsilon}$	$\sqrt{\left(\frac{m\pi x}{a}\right)^2 + \left(\frac{n\pi y}{b}\right)^2} = 2\pi f_c\sqrt{\mu\varepsilon}$

The dominant mode in rectangular waveguide is the TE_{10} mode, which has a cut-off frequency given as

$$f_{c_{10}} = \frac{1}{2a\sqrt{\mu\varepsilon}} = \frac{c}{2a} \tag{30.21}$$

The concept of cut-off frequency is further illustrated by Figure 30.5, a β/k diagram for a lossless WR-90 waveguide (note that in the lossless case, the propagation constant will be equal to $j\beta$). It is apparent that higher-order modes may propagate as the operating frequency increases. At the cut-off frequency, β is zero because the guided wavelength is infinity. At high frequencies, the ratio β/k approaches one.

A number of variations of the rectangular waveguide are available, including single and double-ridged waveguide, which are desirable because of increased bandwidth. However, closed solutions for the fields in these structures do not exist and numerical techniques must be used to solve for the field distributions, as well as essential design information such as guided wavelength and characteristic impedance. In addition, losses are typically higher than standard waveguide.

Circular waveguides are also used in some applications, although not nearly as often as rectangular geometry guides. Closed form solutions for the fields in a circular, perfectly conducting waveguide with an inside diameter of $2a$ are given in Table 30.3. Note that these equations use a standard cylindrical coordinate system with ρ as the radial distance from the z-axis and ϕ as the angular distance measured from the y-axis. The axis of the waveguide is aligned along the z-axis. For both the TE_{mn} and TM_{mn} modes any integer value of $n \geq 0$ is allowed, and $J_n(x)$ and $J_n'(x)$ are Bessel functions of order n and its first derivative. As with rectangular waveguides, only certain values of h are allowed. For the TE_{mn} modes, the allowed values of the modal eigenvalues must satisfy the roots of $J_n'(h_{mn}a) = 0$, where m signifies the root number and may range from one to infinity with $m = 1$, the smallest root. Similarly, for the TM_{nm} modes, the values of the modal eigenvalues are the solutions of $J_n(h_{mn}a) = 0$. The dominant mode in circular waveguides is the TE_{11} mode, with a cut-off frequency given by

$$f_{c_{11}} = \frac{0.293}{a\sqrt{\mu\varepsilon}} \tag{30.22}$$

TABLE 30.3 Field Components for Circular Waveguide

	TE	TM
E_z	0	$E_0 J_n(h_{nm}\rho)\cos(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$
H_z	$H_0 J_n(h_{nm}\rho)\cos(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$	0
E_ρ	$H_0 \frac{j\omega\mu n}{h_{nm}^2\rho} J_n(h_{nm}\rho)\sin(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$	$-E_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho)\cos(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$
H_ρ	$-H_0 \frac{\gamma_{nm}}{h_{nm}} J_n'(h_{nm}\rho)\cos(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$	$-E_0 \frac{j\omega\varepsilon n}{h_{nm}^2\rho} J_n(h_{nm}\rho)\sin(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$
E_φ	$H_0 \frac{j\omega\mu}{h_{nm}} J_n'(h_{nm}\rho)\cos(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$	$E_0 \frac{\gamma_{nm}}{h_{nm}^2\rho} J_n(h_{nm}\rho)\sin(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$
H_φ	$H_0 \frac{\gamma_{nm}}{h_{nm}^2\rho} J_n(h_{nm}\rho)\sin(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$	$-E_0 \frac{j\omega\varepsilon}{h_{nm}} J_n'(h_{nm}\rho)\cos(n\phi)\,\mathrm{e}^{-\gamma_{nm}z}$

TABLE 30.4 Cut-off Frequencies for Several Lower Order Waveguide Modes for Circular Waveguide

f_c/f_{c10}	Modes
1.0	TE_{11}
1.307	TM_{01}
1.66	TE_{21}
2.083	TE_{01}, TM_{11}
2.283	TE_{31}
2.791	TE_{21}
2.89	TE_{41}
3.0	TE_{12}

Frequencies have been normalized to the cut-off frequency of the TE_{10} mode.

Table 30.4 shows the cut-off frequencies for several of the lowest order modes, referenced to the cut-off frequency of the dominant mode.

30.3 Planar Guiding Structures

Planar guiding structures are composed of a comparatively thin dielectric substrate with metallization on one or both planes. By controlling the dimensions of the metallization, a variety of passive components, transmission lines, and matching circuits can be constructed using photolithography and etching or plating techniques. Further, active devices are readily integrated into planar guiding structures. This provides a low-cost and compact way of realizing complicated microwave and millimeter-wave circuits. Microwave integrated circuits (MICs) and monolithic microwave integrated circuits (MMICs) based on this concept are commonly available.

A variety of planar transmission lines have been demonstrated, including microstrip, coplanar waveguide (CPW), slotline, and coplanar stripline (CPS). The cross section of each of these planar transmission lines is shown in Figure 30.7a through d. Once the dielectric substrate is chosen, characteristics of these transmission lines are controlled by the width of the conductors and/or gaps on the top planes of the geometry. Of these, microstrip is by far the most commonly used planar transmission line. CPW is also often used, particularly at millimeter-wave frequencies. Slotline and CPS use is fairly rare in microwave

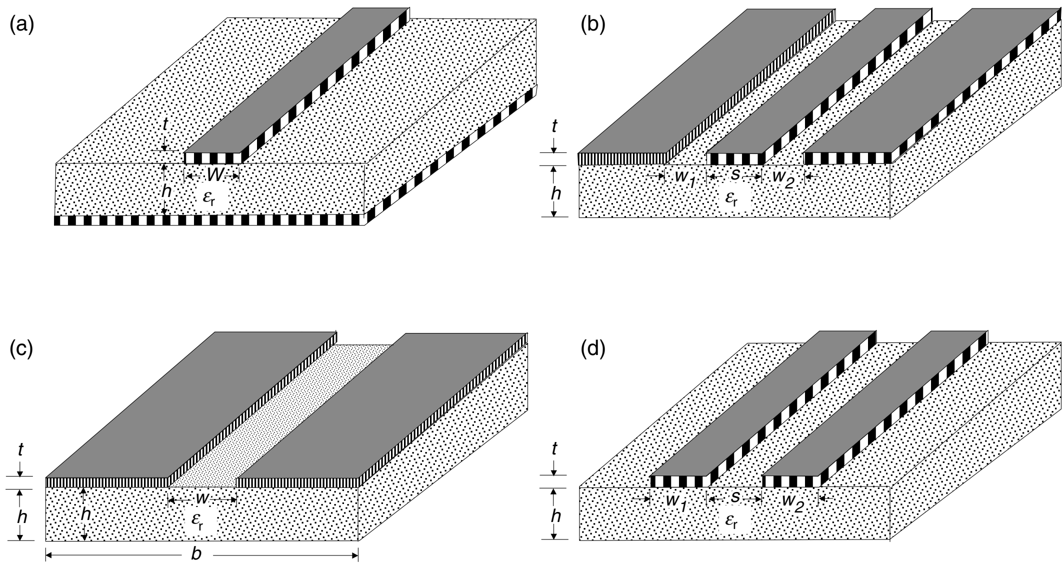


FIGURE 30.7 Cross section of four of the most popular types of planar guiding structures: (a) microstrip, (b) coplanar waveguide, (c) slotline, and (d) coplanar stripline.

and millimeter-wave circuits, but they are sometimes used in transitions and baluns. In this section, we will describe the basic properties of planar transmission lines. Because of its prevalence, microstrip will be described in detail and closed form expressions for the design of microstrip will be given. Readers are referred to [1] for a good general source of information on microstrip and slotline type transmission line structures.

30.3.1 Microstrip

As seen in Figure 30.7a, the simplest form of microstrip consists of a single conductor on a grounded dielectric slab. Microstrip is the most common type of planar transmission line used in microwave and millimeter-wave circuits, with a great deal of design data freely available. A broad range of passive components may be designed with microstrip, including filters, resonators, diplexers, distribution networks, and matching components. In addition, three terminal active components can be integrated by using vias to ground. However, this may introduce considerable inductances at high frequencies, which will reduce the realized gain of active circuits using microstrip.

The fundamental mode of propagation for this type of planar waveguide is often referred to as quasi-TEM, because of its close resemblance to pure TEM modes. In fact, noting that the majority of the power is confined in the region bounded by the width of the microstrip, the basic characteristics of microstrip are quite similar to the parallel-strip transmission line of Figure 30.1a. Because of the presence of the air-dielectric interface, it is not a true TEM mode. The use of the dielectric between the ground and top conductor confines the majority of the fields in this region, but some energy may radiate from the structures. Using a high-permittivity substrate and shielding the structure helps to minimize this factor. Microstrip is capable of carrying moderate power levels (a 50-Ω microstrip line on 25-mil alumina can handle several kW of power depending on conductor metallization), is broadband, and it is possible to realize a variety of circuit topologies, both active and passive.

To design the basic microstrip line, it is necessary to be able to determine characteristic impedance and effective permittivity, preferably as a function of frequency. A wide variety of approximations have been presented in the literature, with most techniques using a quasi-static approximation for the characteristic impedance, Z_0, at low frequencies and then a dispersion model for the characteristic impedance as a function of frequency, $Z_0(f)$, in terms of Z_0. One fairly accurate and simple model that is commonly used to obtain Z_0 and the effective permittivity, ε_{re}, neglecting the effect of conductor thickness is given as [2]:

$$Z_0 = \frac{\eta}{2\pi\sqrt{\varepsilon_{re}}} \ln\left(\frac{8h}{W} + 0.25\frac{W}{h}\right) \quad \text{for } \left(\frac{W}{h} \leq 1\right) \tag{30.23}$$

$$Z_0 = \frac{\eta}{\sqrt{\varepsilon_{re}}} \left\{\frac{W}{h} + 1.393 + 0.667 \ln\left(\frac{W}{h} + 1.444\right)\right\}^{-1} \quad \text{for } \left(\frac{W}{h} \geq 1\right) \tag{30.24}$$

Note that η is 120π-Ω, by definition. The effective permittivity is given as

$$\varepsilon_{re} = \frac{\varepsilon_r + 1}{2} + \frac{\varepsilon_r - 1}{2} F(W/h) \tag{30.25}$$

$$F(W/h) = (1 + 12h/W)^{-1/2} + 0.04(1 - W/h)^2 \quad \text{for } \left(\frac{W}{h} \leq 1\right)$$

$$F(W/h) = (1 + 12h/W)^{-1/2} \quad \text{for } \left(\frac{W}{h} \geq 1\right)$$

With these equations, one can determine the characteristic impedance in terms of the geometry. For desired characteristic impedance, the line width can be determined from

$$W/h = \frac{8\exp(A)}{\exp(2A) - 2} \quad \text{for } A > 1.52 \tag{30.26}$$

$$W/h = \frac{2}{\pi}\left\{B - 1 - \ln(2B - 1) + \frac{\varepsilon_r - 1}{2\varepsilon_r}\left[\ln(B - 1) + 0.39 - \frac{0.61}{\varepsilon_r}\right]\right\} \text{ for } A > 1.52 \tag{30.27}$$

where

$$A = \frac{Z_0}{60}\left\{\frac{\varepsilon_r + 1}{2}\right\}^{1/2} + \frac{\varepsilon_r - 1}{\varepsilon_r + 1}\left\{0.23 + \frac{0.11}{\varepsilon_r}\right\}$$

$$B = \frac{60\pi^2}{Z_0\sqrt{\varepsilon_r}}$$

Once Z_0 and ε_{re} have been determined, effects of dispersion may also be determined using expressions from Hammerstad and Jensen [3] for $Z_0(f)$ and Kobayashi [4] for $\varepsilon_{re}(f)$. To illustrate the effects of dispersion, the characteristic impedance and effective permittivity of several microstrip lines on various substrates are plotted in Figure 30.8a and b, using the formulas from the previously mentioned papers. The substrates indicated by the solid ($\varepsilon_r = 2.33$, $h = 31$ mil, $W = 90$ mil) and dashed ($\varepsilon_r = 10.2$, $h = 25$ mil, W $= 23$ mil) lines in these figures are typical for those that might be used in a hybrid circuit at microwave frequencies. We can see in Figure 30.8a that the characteristic impedance is fairly flat until ~10 GHz, above which it may be necessary to consider the effects of dispersion for accurate design. The third line in the figure is an alumina substrate ($\varepsilon_r = 9$, $h = 2.464$ mil, $W = 2.5$ mil) on a thin substrate. The characteristic impedance is flat until about 70 GHz, indicating that this thin substrate is useful at higher frequency operation. Figure 30.8b shows the effective permittivity as a function of frequency. Frequency variation for this parameter is more dramatic. However, it must be remembered that guided wavelength is inversely proportional to the square root of the effective permittivity. Therefore, variation in electrical length will be less pronounced than the plot suggests.

In addition to dispersion, higher frequency operation is complicated by a number of issues, including decreased Q-factor, radiation losses, surface wave losses, and higher-order mode propagation. The designer must be aware of the limitations of both the substrate and the characteristic impedance of the

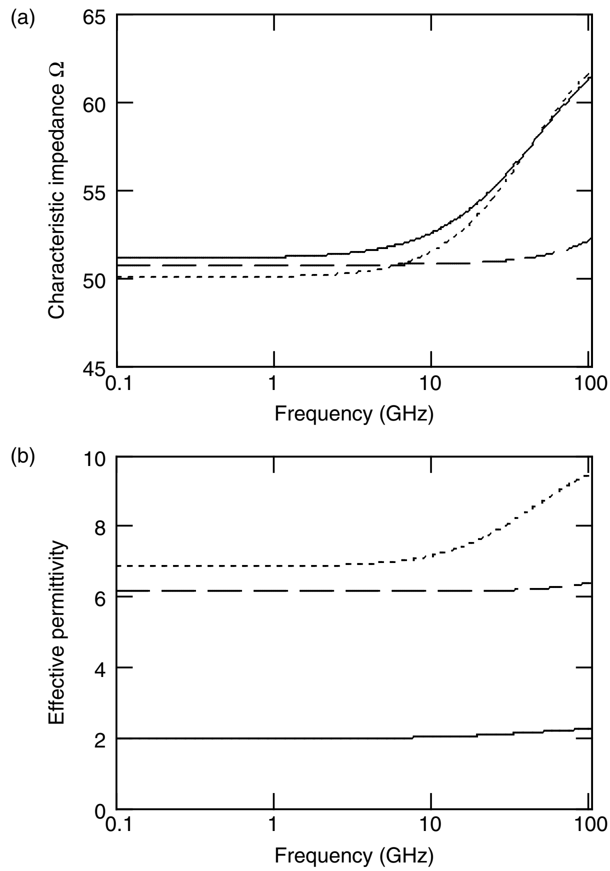


FIGURE 30.8 Dispersion characteristics of 50 Ω line on three substrates (solid line is $\varepsilon_r = 2.33$, $h = 31$ mil, $W = 90$ mil; dotted line is $\varepsilon_r = 10.2$, $h = 25$ mil, $W = 23$ mil; dashed line is $\varepsilon_r = 9$, $h = 2.464$ mil, $W = 2.5$ mil). Shown in (a), the impedance changes significantly at high frequencies for the thicker substrates as does the effective permittivity shown in (b).

lines. In terms of the substrate, considerable amount of energy can couple between the desired quasi-TEM mode of the microstrip and the lowest order surface wave mode of the substrate. In terms of the substrate thickness and permittivity, an approximation for determining the frequency where this coupling becomes significant is given by the following expression [5]:

$$f_T = \frac{150}{\pi h}\sqrt{\frac{2}{\varepsilon_r - 1}\arctan(\varepsilon_r)} \tag{30.28}$$

Note that f_T is in gigahertz and h is in millimeters. In addition to the quasi-TEM mode, microstrip will propagate undesired higher order TE- and TM-type modes with cut-off frequency roughly determined by the cross section of the microstrip. The excitation of the first mode is approximately given by the following expression [5]:

$$f_c = \frac{300}{\sqrt{\varepsilon_r}(2W + 0.8h)} \tag{30.29}$$

Again, note that f_c is in gigahertz, and h and W are both in millimeters. This expression is useful in determining the lowest impedance that may be reliably used for a given substrate and operating frequency.

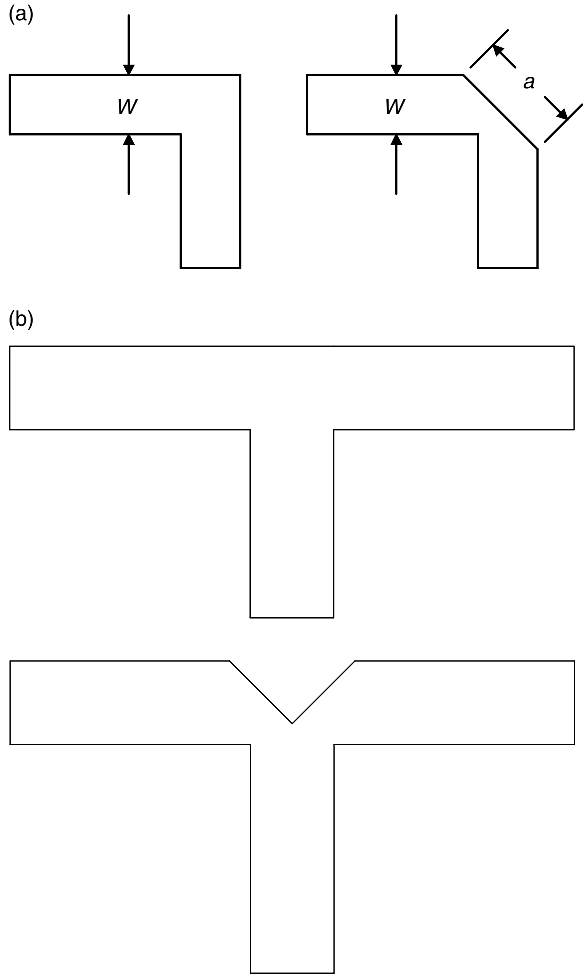


FIGURE 30.9 Two common microstrip discontinuities encountered in layout: (a) the microstrip bend and (b) the T-junction.

As a rule of thumb, the maximum operating frequency should be chosen somewhat lower. A good choice for maximum frequency may be 90% of this value or lower.

A variety of techniques have also been developed to minimize or characterize the effects of discontinuities in microstrip circuits, a variety of which are shown in Figure 30.9a and b including a microstrip bend and a T-junction. Another common effect is the fringing capacitance found at impedance steps or open-circuited microstrip stubs.

The microstrip bend allows flexibility in microstrip circuit layouts and may be at an arbitrary angle with different line widths at either end. However, by far, the most common is the 90° bend with equal widths at either end, shown on the left of Figure 30.8a. Owing to the geometry of the bend, excess capacitance is formed causing a discontinuity. A variety of techniques have been used to reduce the discontinuity by eliminating a sufficient amount of capacitance, including the mitered bend shown on the right. Note that another way of reducing this effect is to use a curved microstrip line with sufficiently large radius. A second type of discontinuity commonly encountered by necessity in layouts is the T-junction, shown in Figure 30.9b, which is formed at a junction of two lines. As with the bend, excess capacitance is formed, degrading performance. The mitered T-junction on the right is used to reduce this problem. Again, a variety of other simple techniques have also been developed.

Fringing capacitance will be present with microstrip open-circuited stubs and at impedance steps. With the open-circuited stub, this causes the electrical length of the structure to be somewhat longer. For an impedance step, the lower impedance line will also appear to be electrically longer. The simplest way of compensating for this problem is by modeling the capacitance and effective length of the fringing fields. Again, a variety of simple models have been developed to perform this task, most of them based on quasi-static approximations. A commonly used expression for the length extension of an open-end based on empirical data is given by [6]

$$\frac{\Delta l_{\mathrm{oc}}}{h} = 0.412 \frac{\varepsilon_{\mathrm{re}} + 0.3}{\varepsilon_{\mathrm{re}} - 0.258} \left[\frac{W/h + 0.264}{W/h + 0.8} \right] \tag{30.30}$$

This expression is reported to yield relatively accurate results for substrates with permittivity in the range of 2–50 but is not as accurate for wide microstrip lines. For the impedance step, a first-order approximation for determining the excess length of the impedance step is to multiply the open-end extension, $\Delta l_{\mathrm{oc}}/h$, by an appropriate factor to obtain a useful value, that is, $\Delta l_{\mathrm{step}}/h \approx \Delta l_{\mathrm{oc}}(w_1/w_2 - 1)/h$.

Because of the prevalence of microstrip, modern microwave CAD tools typically have extensive libraries for microstrip components, including discontinuities effects.

30.3.2 Coplanar Waveguide

Coplanar waveguide (CPW), shown in Figure 30.7b, consists of a signal line and two ground planes on a dielectric slab with metallization on one side. For a given substrate, characteristic impedance is determined by the signal line width, s, and the two gaps, w_1 and w_2. This structure often demonstrates better dispersion characteristics than microstrip. In addition, three terminal devices are easily integrated into this uniplanar transmission line that requires no vias for grounding. For this reason, inductive parasitics are lower than microstrip, making CPW a good choice for high-frequency operation where this is a primary design concern. For this reason, a significant portion of MMICs operating at W-band and above are realized in CPW. CPW does find some application at lower frequencies as well due to its topside ground access and low coupling characteristics. For instance, backside vias are difficult to process on SiC and it may be more cost-effective to use CPW. Mixers and switches are sometimes realized in CPW because of simplicity in realizing baluns and low coupling due to the boundary condition imposed by the topside ground.

The three-conductor line shown in Figure 30.7b supports two fundamental modes, including the desired CPW-mode and an undesired coupled slotline mode if the two ground planes separating the signal line are not kept at the same potential. For this reason, wires or metal strips referred to as *airbridges* are placed at discontinuities where mode conversion may occur.

Packaging may be a problem for this type of structure, because the bottom plane of the dielectric may come in close proximity to other materials, causing perturbations of the transmission line characteristics. In practice, this is remedied by using *grounded* or *conductor-backed* CPW (CB-CPW) where a ground plane is placed on the backside for electrical isolation. At high frequencies, this may present a problem with additional losses through coupling to the parallel-plate waveguide mode. These losses can be minimized by placing vias in the region around the transmission line. In addition, the topside ground planes must be of finite dimensions. At higher frequencies, the ground metallization will resonate. This resonance is analogous to the operation of a microstrip patch antenna formed by a resonant dimension metallization on the topside of a conductor-backed substrate. This can also be eliminated by placing backside vias in the CPW ground planes. In cases where CB-CPW is used without backside vias, analysis should be used to determine the maximum ground dimensions for a given frequency of operation.

Although CPW was first proposed by Wen [7] in 1969, acceptance of CPW has been much slower than microstrip. For this reason, simple and reliable models for CPW are not as readily available as for microstrip. A compilation of some of the more useful data can be found in [8].

30.3.3 Slotline and Coplanar Stripline

Two other types of planar transmission lines are slotline and CPS. These structures are used less often than either microstrip or CPW, but do find some applications. Both of these structures consist of a dielectric slab with metallization on one side. Slotline has a slot of width w etched into the ground plane. CPS consists of two metal strips of width w_1 and w_2 separated by a distance s on the dielectric slab. Owing to their geometry, both these structures are balanced transmission line structures and are useful in balanced circuits such as mixers and modulators. Only limited design information is available for both these types of transmission line.

The slotline mode is non-TEM and is almost entirely TE. However, no cut-off frequency exists as with the waveguide TE modes discussed in Section 30.2. Microwave circuits designed solely in slotline are seldom used. However, slotline is sometimes used in conjunction with other transmission line types, such as microstrip or CPW, for increased versatility. Examples of these include filters, hybrids, and resonators. In addition, slotline is sometimes used in planar antennas, such as the slot antenna or some kinds of multilayer patch antennas.

The CPS transmission line has two conductors on the top plane of the circuit, allowing series or shunt elements to be readily integrated into CPS circuits. CPS is often used in electro-optic circuits, such as optic traveling wave modulators, as well as in high-speed digital circuits. Owing to its balanced nature, CPS also makes an ideal feed for printed dipoles. Difficulties (or benefits, depending on the application) with CPS include high characteristic impedances.

References

1. K.C. Gupta, R. Garg, I. Bahl, R. Bhartia, *Microstrip Lines and Slotlines*, Artech House, Inc., Norwood, MA, 1996.
2. E. Hammerstad, "Equations for microstrip circuit design," *Proc. European Microwave Conf.*, 1975, pp. 268–272.
3. E. Hammerstad and O. Jensen, "Accurate models for microstrip computer-aided design," *IEEE MTT-S Int. Microwave Symp. Dig.*, 1980, pp. 407–409.
4. M. Kobayashi, "A dispersion formula satisfying recent requirements in microstrip CAD," *IEEE Trans.*, Vol. MTT-36, August 1988, pp. 1246–1250.
5. G.D. Vendelin, "Limitations on stripline Q," *Microwave J.*, Vol. 13, May 1970, pp. 63–69.
6. R. Garg and I.J. Bahl, "Microstrip discontinuities," *Int. J. Electron.*, Vol. 45, July 1978, pp. 81–87.
7. C.P. Wen, "Coplanar Waveguide: a surface strip transmission line suitable for nonreciprocal gryomagnetic device applications," *IEEE Trans. Microwave Theory Tech.*, Vol. 112, December 1969, pp. 1087–1090.
8. R.N. Simons, *Coplanar Waveguide Circuits, Components and Systems*, Wiley Inter-Science, New York, NY, 2001.

31

The Effects of Multipath Fading in Wireless Communication Systems

Wayne Stark
University of Michigan

31.1 Introduction

The performance of a wireless communication system is heavily dependent on the channel over which the transmitted signal propagates. Typically in a wireless communication system the channel consists of multiple paths between the transmitter and receiver with different attenuation and delay. The paths have different attenuation and delays because of the different distances between transmitter and receiver along different paths. For certain transmitted signals the entire transmitted signal may experience a deep fade (large attenuation) due to destructive multipath cancellation. The multipath signals may also add constructively giving a larger amplitude. In addition to the multipath effect of the channel on the transmitted signal, there are other effects on the transmitted signal due to the channel. One of these is distance related and is called the propagation loss. The larger the distance between the transmitter and receiver the smaller the received power. Another effect is known as shadowing. Shadowing occurs due to building and other obstacles obstructing the line of sight path between the transmitter and receiver. This causes the received signal amplitude to vary as the receiver moves out from behind buildings or moves behind buildings. In this chapter we examine models of fading channels and methods of mitigating the degradation in performance due to fading. We first discuss in detail models for the multipath fading effects of wireless channels. Then briefly we discuss the models for propagation loss as a function of distance and shadowing. Next we show an example how diversity in receiving over multiple, independent faded paths can significantly improve performance. We conclude by discussing the fundamental limits on reliable communication in the presence of fading.

31.2 Multipath Fading

In this section, we discuss the effects of multiple paths between the transmitter and receiver. These effects depend not only on the delays and amplitudes of the paths but also on the transmitted signal. We give examples of frequency selective fading and time selective fading.

Consider a sinusoidal signal transmitted over a multipath channel. If there are two paths between the transmitter and receiver with the same delay and amplitude but opposite phase (180° phase shift) the channel will cause the received signal amplitude to be zero. This can be viewed as destructive interference. However if there is no phase shift then the received signal amplitude will be twice as large as the signal amplitude on each of the individual paths. This is the case of constructive interference.

In a digital communication system, data is modulated onto a carrier. The data modulation causes variations in the amplitude and phase of the carrier. These variations occur at a rate proportional to the bandwidth of the modulating signal. As an example, if the data rate is on the order of 25 kbps and the carrier frequency is about 900 MHz the amplitude and phase of the carrier will change at a rate on the order of 25 kHz. Equivalently, the envelope and phase of the carrier might change significantly every $1/25\,\text{kHz} = 0.04\,\text{ms} = 40\,\mu\text{s}$. If this signal is transmitted over a channel with two paths with differential delay of 1 μs, say, then the modulation part of the signal would not differ significantly on these two paths since in a 1 μs period the modulation does not change significantly. However, if this signal was received on two paths with a differential delay of 40 μs then the received signal on these two paths would be significantly different, because, in 40 μs, the modulated part of the signal would have changed. If the data rate of the signal increased to 250 kbps then the modulated signal would change significantly in a 4 μs time frame and thus the effect of multipath would be different.

Thus, the type of fading depends on various parameters of the channel and the transmitted signal. Fading can be considered as filtering operation on the transmitted signal. The filter characteristics are, in general, time varying due to the motion of the transmitter/receiver or objects in the environment. The faster the motion the faster the change in the filter characteristics. In addition, the filter characteristics cause the channel to have time dispersion, which is equivalent to frequency selectivity. In other words, different frequencies may have different responses. Fading channels are typically characterized in the following ways:

1. *Frequency Selective Fading:* If the transfer function of the filter has significant variations within the frequency band of the transmitted signal then the fading is called frequency selective.
2. *Time Selective Fading:* If the fading changes relatively quickly (compared to the duration of a data bit) then the fading is said to be time selective.

If the channel is both time and frequency selective then it is said to be doubly selective.

To illustrate these types of fading we consider some special cases. Consider a simple model for fading where there are a finite number, k, of paths from the transmitter to the receiver and these paths are not time dependent. The transmitted signal is denoted by $s(t)$. The signal can be represented as a base band signal modulated onto a carrier as

$$s(t) = Re[s_0(t) \exp\{j2\pi f_c t\}]$$

where f_c is the carrier frequency and $s_0(t)$ is the baseband signal or the envelope of the signal $s(t)$. The paths between the transmitter and receiver have delays τ_k and amplitudes α_k and phases ϕ_k. The impulse response of the channel is

$$h(t) = \sum_{k=1}^{M} \alpha_k e^{j\phi_k} \delta(t - \tau_k)$$

and the transfer function is

$$H(f) = \sum_{k=1}^{M} \alpha_k \exp\{j\phi_k - j2\pi f \tau_k\}.$$

The received signal can thus be expressed as

$$r(t) = Re\left[\sum_k \alpha_k s_0(t - \tau_k) \exp\{j2\pi f_c(t - \tau_k) + j\phi_k\}\right].$$

The baseband received signal is given by

$$r_0(t) = \sum_k \alpha_k s_0(t - \tau_k) \exp\{j\phi_k - j2\pi f_c \tau_k\}.$$

To understand the effects of multipath, we will consider a couple of different examples for a time nonselective channel.

31.2.1 Frequency/Time Nonselective (Flat) Fading

First we consider a frequency and time nonselective fading model. In this case the multipath components are assumed to have independent phases. If we let W denote the bandwidth of the transmitted signal then the envelope of the signal does not change significantly in time smaller than $1/W$. Thus if the maximum delay satisfies $\tau_{\max} \ll 1/W$, that is,

$$\frac{1}{f_c} \ll \tau_k \ll T = W^{-1}$$

then, $s_0(t - \tau_k) \approx s_0(t)$. In this case,

$$\begin{aligned} r_0(t) &= s_0(t)\left(\sum_k \alpha_k \exp\{j\theta_k\}\right) \\ &= X s_0(t) \end{aligned}$$

where, $\theta_k = \phi_k - 2\pi f_c \tau_k$. The factor $X = \sum_k \alpha_k \exp\{j\theta_k\}$ by which the signal is attenuated/phase shifted is usually modeled by a complex Gaussian distributed random variable. The magnitude of X is a Rayleigh-distributed random variable. The phase of X is uniformly distributed. The fading occurs because of the random phases sometimes adding destructively and sometimes adding constructively. Thus, for narrow enough signal bandwidths ($\tau_k \ll W^{-1}$) the multipath results in an amplitude attenuation by a Rayleigh-distributed random variable. It is important to note that the transmitted signal in this example has not been distorted. The only effect on the transmitted signal is an amplitude and phase change. This will not be true in a frequency selective channel.

31.2.2 Frequency Selective/Time Nonselective Fading

Now consider the case where the bandwidth of the modulating signal $s_0(t)$ is W and the delays satisfy

$$\tau_k \gg T = W^{-1}.$$

In this case we say the channel exhibits frequency selective fading. The baseband received signal for M paths is

$$r_0(t) = \alpha_1 e^{j\theta_1} s_0(t - \tau_1) + \cdots + \alpha_M s_0(t - \tau_M) e^{j\theta M}.$$

Consider, for simplicity, the case $M = 2$. Assume that the receiver is synchronized to the first path so that we can assume $\tau_1 = \phi_1 = \theta_1 = 0$. Then the transfer function can be written as

$$H(f) = 1 + \alpha_2 \exp\{j\theta_2 - j2\pi f \tau_2\}.$$

At frequencies where $2\pi f \tau_2 = \theta_2 + 2n\pi$ or $f = (\theta_2 + 2n\pi)/2\pi\tau_2$ for some integer n the transfer function will be $H(f) = 1 + \alpha_2$. If $\alpha_2 > 0$ then the amplitude of the received signal will be larger because of the second path. This is called constructive interference. At frequencies where $2\pi f \tau_2 = \theta_2 + (2n+1)\pi$ or $f = (\theta_2 + (2n+1)\pi)/2\pi\tau_2$ the transfer will be $H(f) = 1 - \alpha_2$. Again, for $\alpha_2 > 0$ the amplitude of the received signal will be smaller due to the second path. This is called destructive interference. The frequency range between successive nulls (destructive interference) is $1/\tau$. Thus if $\tau \gg \frac{1}{W}, \frac{1}{\tau} \ll W$ there will be multiple nulls in the spectrum of the received signal. In Figure 31.1 we show the transfer function of a multipath channel with two equal strength paths with differential delay of 1 μs. In Figure 31.2, we show the transfer function of a channel with eight equal strength paths with delays from 0 to 7 μs. The frequency selectivity of the channel is seen by the fact that the transfer function varies as a function of frequency. Narrowband systems might have spectrum narrow enough so that the whole signal band experiences a

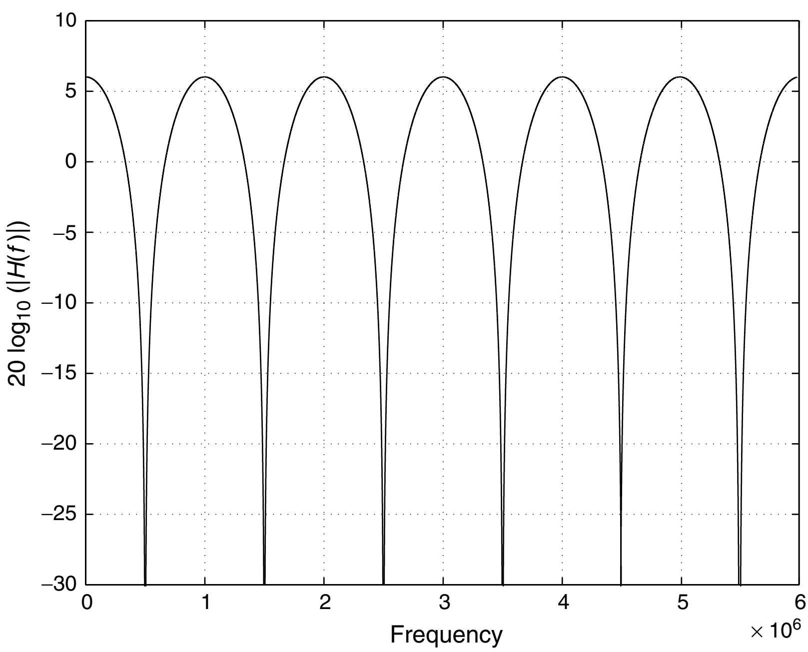


FIGURE 31.1 Transfer function of multipath channel with two equal strength paths and relative delay of 1 μs.

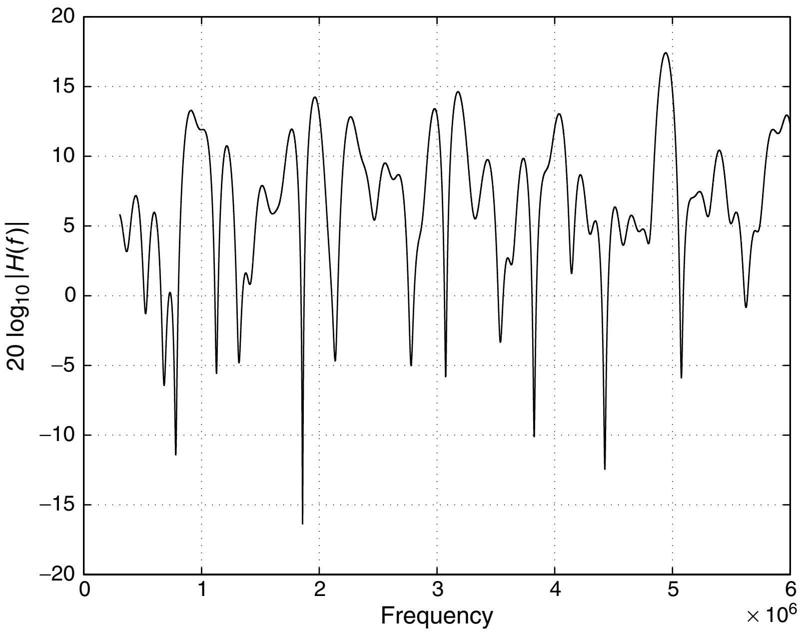


FIGURE 31.2 Transfer function of multipath channel with eight equal strength paths.

deep fade while in wideband systems the transfer function of the channel varies within the band of the transmitted signal and thus the channel causes distortion of the transmitted signal.

31.2.3 Time Selectivity

The dual concept to frequency selectivity is time selectivity. In this case the path strength is changing as a function of time (e.g., due to vehicle motion) and the envelope of the received signal (as the vehicle moves) undergoes time-dependent fading. Usually, the paths lengths change with time due to motion of the transmitter or receiver. Here, we assume that the motion is slow enough relative to the symbol duration so that the $\alpha_k(t)$ and $\phi_k(t)$ are constants over the duration of a symbol but vary from one symbol to the next. In this model, the transmitted signal is simply attenuated by a slowly varying random variable. This is called a flat fading model or frequency and time nonselective fading.

The channel model for a time-selective (but frequency nonselective fading) is that of a time varying impulse response:

$$h(t; t - \beta) = \alpha_k(t)e^{j\theta(t)}\delta(t - \beta - \tau(t))$$

where $\tau(t)$ is the time varying delay between the transmitter and receiver. The output $r_0(t)$ of the channel is related to the input $s_0(t)$ via

$$\begin{aligned} r_0(t) &= \int_{-\infty}^{\infty} h(t; t - \beta)s_0(\beta)\mathrm{d}\beta \\ &= \int_{-\infty}^{\infty} \alpha_k(t)e^{j\theta(t)}\delta(t - \beta - \tau(t))s_0(\beta)\mathrm{d}\beta \\ &= \alpha_k(t)e^{j\theta(t)}s_0(t - \tau(t)). \end{aligned}$$

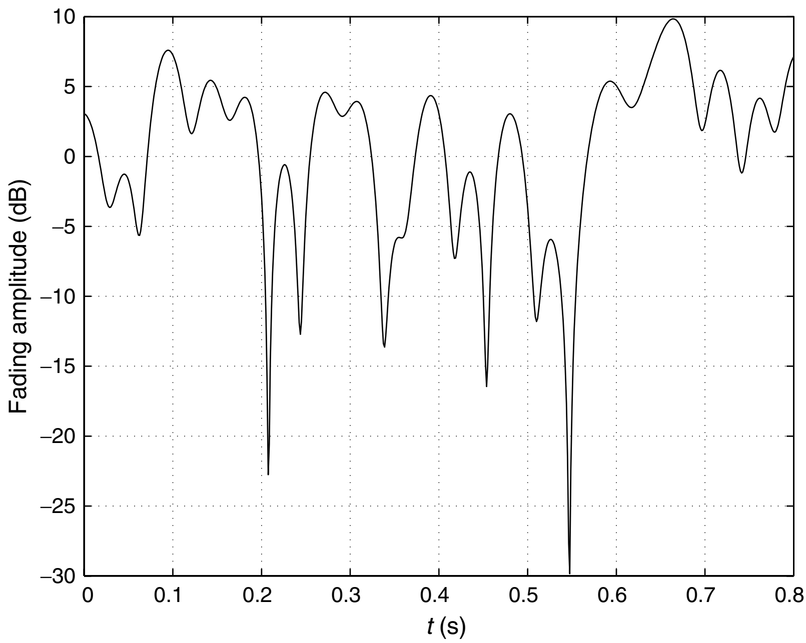


FIGURE 31.3 Received signal strength as a function of time for vehicle velocity 10 mph.

Because the impulse response is time-varying the fading at different time instances is correlated if the time instances are very close and uncorrelated if they are very far apart. Consider the simple case of a sinusoidal at frequency f_c as the signal transmitted. In this case the baseband component of the transmitted signal, $s_0(t)$, is a constant DC term. However the output of the channel is given by

$$r_0(t) = \alpha_k(t)e^{j\theta(t)}s_0(t).$$

The frequency content of the baseband representation of the received signal is no longer just a DC component but has components at other frequencies due to the time varying nature of α and θ. If we consider just a single direct path between the transmitter and receiver and assume that the receiver is moving away from the transmitter then because of the motion there will be a Doppler shift in the received spectrum. In other words, the received frequency will be shifted down in frequency. Similarly if the receiver is moving toward the transmitter there will be a shift up in the frequency of the received signal. Because there can be paths between the transmitter and receiver that are direct and paths that are reflected, some paths will be shifted up in frequency and some paths will be shifted down in frequency. The overall received signal will be spread out in the frequency domain due to these different frequency shifts on different paths. The spread in the spectrum of the transmitted signal is known as the Doppler spread. If the data duration is much shorter than the time variation of the fading process then the fading can be considered a constant or a slowly changing random process. In Figure 31.3, we plot the fading amplitude for a single path as a function of time for a mobile traveling at 10 miles/h (mph). In Figure 31.4, a similar plot is done for a mobile at 30 mph. It is clear that the faster a mobile is moving the more quickly the fading amplitude varies. Fading amplitude variations with time can be compensated for by power control at low vehicle velocities. At high velocities the changes in amplitude can be mitigated by proper use of error control coding.

31.3 General Model

In this section, we describe a general model for fading channels and discuss the relevant parameters that characterize a fading channel. The most widely used general model for fading channels is the wide-sense

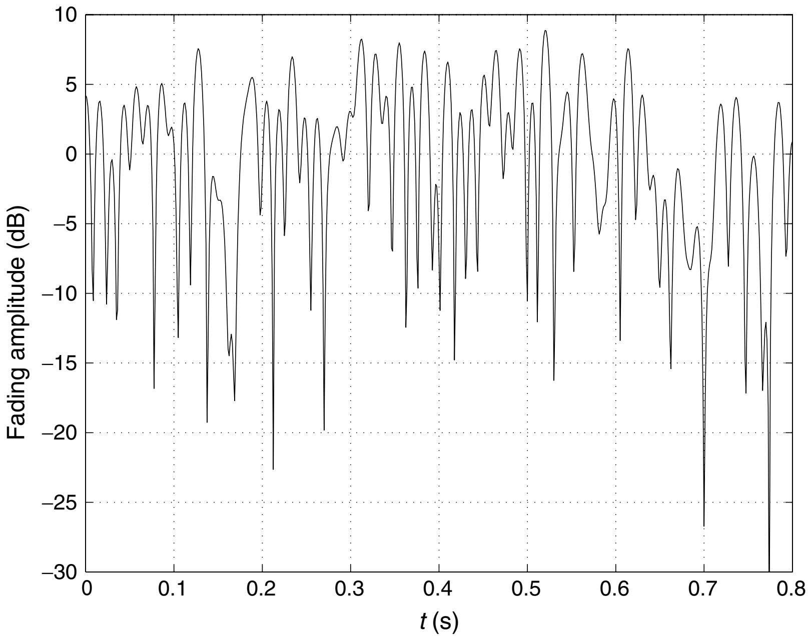


FIGURE 31.4 Received signal strength as a function of time for vehicle velocity 30 mph.

stationary, uncorrelated scattering (WSSUS) fading model. In this model, the received signal is modeled as a time-varying filter operation on the transmitted signal. In other words,

$$r_0(t) = \int_{-\infty}^{\infty} h(t; t - \alpha) s_0(\alpha) \mathrm{d}\alpha$$

where $h(t; t - \tau)$ is the response due to an impulse at time τ and is modeled as a zero mean complex Gaussian random process. Note that it depends not only on the time difference between the output and the input but also on the time directly. The first variable in h accounts for the time-varying nature of the channel while the second variable accounts for the delay between the input and output. This is the result of the assumption that there are a large number of (possibly time varying) paths at a given delay with independent phases. If there is no direct (unfaded) path then the impulse response will have zero mean. In this case the channel is known as a Rayleigh faded channel. If there is a (strong) direct path between the transmitter and receiver then the filter $h(t, \tau)$ will have nonzero mean. This case is called a Rician faded channel. In the following we will assume the mean of the channel is zero.

The assumption for WSSUS is that the impulse response, $h(t, \tau)$, is uncorrelated for different delays and the correlation at different times depends only on the time difference. Mathematically, we write the correlation of the impulse response at different delays and times as an expectation;

$$E[h(t; \tau_1) h^*(t + \Delta t; \tau_2)] = \phi(\tau_1; \Delta t) \delta(\tau_2 - \tau_1)$$

where $E[h(t; \tau_1) h^*(t + \Delta t; \tau_2)]$ denotes the expected (average) value of the impulse response at two different delays and times. The function $\phi(\tau; \Delta t)$ is the intensity delay profile and $\delta(\tau)$ is the usual Dirac delta function. The amount of power received at a given delay τ is $\phi(\tau; 0)$. This is called the intensity delay profile or the delay power spectrum. The mean excess delay, μ_{m} is defined to be the average excess delay above the delay of the first path

$$\mu = \frac{\int_{\tau_{\min}}^{\tau_{\max}} \tau \phi(\tau; 0) \mathrm{d}\tau}{\int_{\tau_{\min}}^{\tau_{\max}} \phi(\tau; 0) \mathrm{d}\tau} - \tau_{\min}.$$

The rms delay spread is defined as

$$s = \left[\frac{\int_{\tau_{\min}}^{\tau_{\max}} (\tau - \mu - \tau_{\min})^2 \, \phi(\tau;0) \mathrm{d}\tau}{\int_{\tau_{\min}}^{\tau_{\max}} \phi(\tau;0) \mathrm{d}\tau} \right]^{1/2}.$$

The largest value $\tau_{\max}$ of τ such that $\phi(\tau;0)$ is nonzero is called the multipath spread of the channel. The importance of the rms delay spread is that it is a good indicator of the performance of a communication system with frequency selective fading. The larger the rms delay spread the more intersymbol interference. In the general model the larger the rms delay the more distortion in the received signal.

Now consider the frequency domain representation of the channel response. The time-varying transfer function of the channel $H(f;t)$ is given by the Fourier transform of the impulse response with respect to the delay variable, that is,

$$H(f;t) = \int_{-\infty}^{\infty} h(t;\tau) \mathrm{e}^{-j2\pi f \tau} \mathrm{d}\tau.$$

Since $h(t;\tau)$ is assumed to be a complex Gaussian random variable, $H(f;t)$ is also a complex Gaussian random process. The correlation $\Phi(f_1, f_2; \Delta t)$ between the transfer function at two different frequencies and two different times is defined as

$$\begin{aligned} \Phi(f_1, f_2; \Delta t) &= E[H(f_1;t) H^*(f_2; t + \Delta t)] \\ &= \int_{-\infty}^{\infty} \phi(\tau; \Delta t) \mathrm{e}^{-j2\pi (f_2 - f_1)\tau} \mathrm{d}\tau. \end{aligned}$$

Thus, the correlation of the channel response between two frequencies (and at two times) for the WSSUS model depends only on the frequency difference. If we let $\Delta t = 0$ then we obtain

$$\Phi(\Delta f; 0) = \int_{-\infty}^{\infty} \phi(\tau;0) \mathrm{e}^{-j2\pi(\Delta f)\tau} \mathrm{d}\tau.$$

As the frequency separation becomes larger the correlation in the response between those two frequencies generally decreases. The smallest frequency separation, B_c, such that the correlation of the response at two frequencies separated by B_c is zero is called the coherence bandwidth of the channel. It is related to the delay spread by

$$B_c \approx \frac{1}{\tau_{\max}}.$$

The rms delay spread and coherence bandwidth are important measures for narrowband channels. The performance of an equalizer for narrow band channels often does not depend on the exact delay power profile but just on the rms delay spread.

Now consider the time-varying nature of the channel. In particular, consider $\Phi(\Delta f; \Delta t)$, which is the correlation between the responses of the channel at two frequencies separated by Δf and at times separated by Δt. For $\Delta f = 0, \Phi(0; \Delta t)$ measures the correlation between two responses separated in time by Δt (at the same frequency). The Fourier transform of $\Phi(0;\gamma)$ gives the Doppler power spectral density

$$S(\lambda) = \int_{-\infty}^{\infty} \Phi(0;\gamma) \mathrm{e}^{-je\pi\lambda\gamma} \mathrm{d}\gamma.$$

The Doppler power spectral density gives the distribution of received power as a function of frequency shift. Since there are many paths coming from different directions and the receiver is moving these paths will experience different frequency shifts.

Example: Consider a situation where a mobile is moving toward a base station with velocity v. If we assume that there are many multipath components that arrive with an angle uniformly distributed over $[0, 2\pi]$ then the Doppler spectral density is given by

$$S(\lambda) = \frac{1}{2\pi f_m}\left[1 - (\lambda/f_m)^2\right]^{-1/2}, \quad 0 \le |\lambda| \le f_m$$

where $f_\mathrm{m} = vf_\mathrm{c}/c$, f_c is the center frequency and c is the speed of light (3×10^8 m/s). For example, a vehicle moving at 100 m/s with 1 GHz center frequency has maximum Doppler shift of 33.3 Hz. A vehicle moving at 30 m/s would have a maximum Doppler shift of 10 Hz. Thus most of the power is either at the carrier frequency +10 Hz or at the carrier frequency −10 Hz. The corresponding autocorrelation function is the inverse Fourier transform and is given by

$$\Phi(0, \gamma) = \int_{-\infty}^{\infty} S(\lambda)e^{j2\pi\lambda\gamma}\,\mathrm{d}\lambda = J_0(2\pi f_\mathrm{m}\gamma).$$

The channel correlation and Doppler spread are illustrated in Figures 31.5 and 31.6 for vehicle velocities of 10 km/h and 100 km/h. From these figures, it is clear that a lower vehicle velocity implies a small spread in the spectrum of the received signal and a larger correlation between the fading at different times. It is often useful for the receiver in a digital communication system to estimate the fading level. The faster the fading level changes the harder it is to estimate. The product of maximum Doppler spread f_m times the data symbol duration T is a useful for determining the difficulty in estimating the channel response. For $f_\mathrm{m}T$ products much smaller than 1 the channel is easy to estimate while for $f_\mathrm{m}T$ much larger than 1 the channel is hard to estimate. Channel estimation can improve the performance of coded systems as shown in the last section of this chapter. The availability of channel information is sometimes called "side information."

If the channel is not time varying (i.e., time invariant) then the response at two different times are perfectly correlated so that $\Phi(0; \Delta t) = 1$. This implies that $S(\lambda) = \delta(f)$. The largest value of λ for which $S(\lambda)$ is nonzero is called the Doppler spread of the channel. It is related to the coherence time T_c, the largest time difference for which the responses are correlated by

$$B_\mathrm{d} = \frac{1}{T_\mathrm{c}}.$$

31.4 GSM Model

The global system for mobile communications (GSM) model was developed in order to compare different coding and modulation techniques. The GSM model is a special case of the general WSSUS model described in the previous section. The model consists of N_p paths, each time varying with different power levels. In Figure 31.7, one example of the delay power profile for a GSM model of an urban environment is shown. In the model each path's time variation is modeled according to a Doppler spread for a uniform angle of arrival spread for the multipath. Thus, the vehicle velocity determines the time selectivity for each path. The power delay profile shown below determines the frequency selectivity of the channel.

In Table 31.1 the parameters for the GSM model are given. The usefulness of this model is that it gives communications engineers a common channel to compare the performance of different designs.

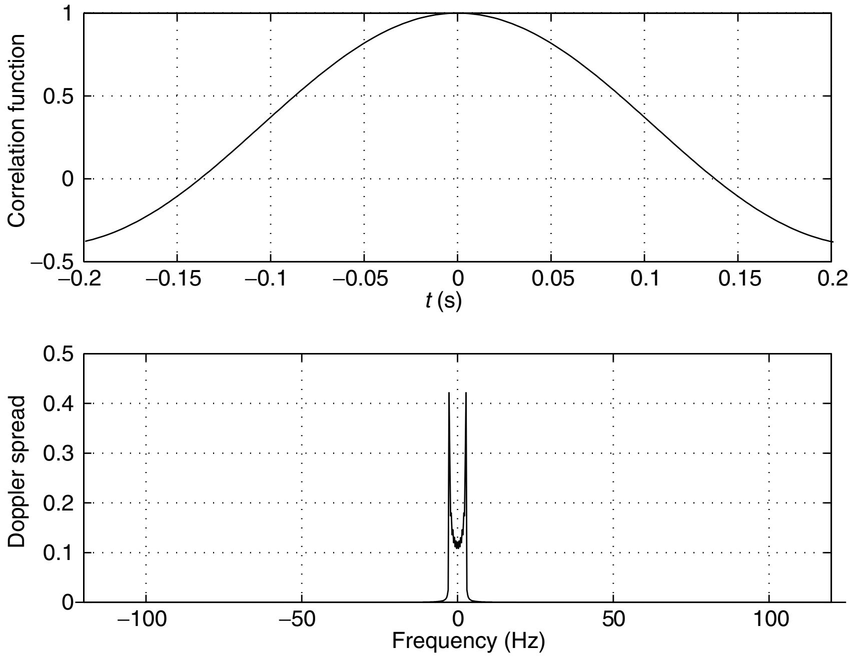


FIGURE 31.5 Channel correlation function and Doppler spread for $f_c = 1$ GHz, $v = 10$ km/h.

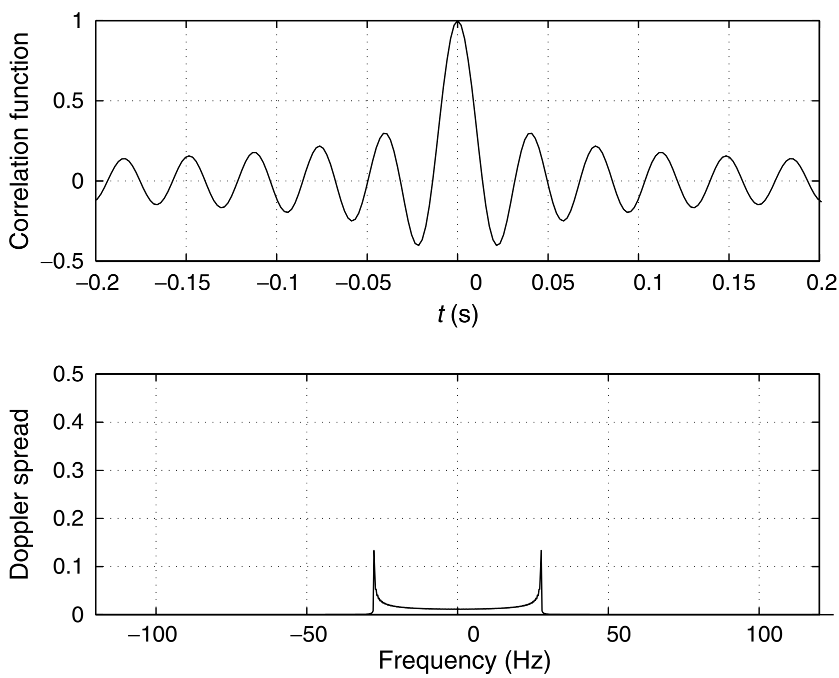


FIGURE 31.6 Channel correlation function and Doppler spread for $f_c = 1$ GHz, $v = 100$ km/h.

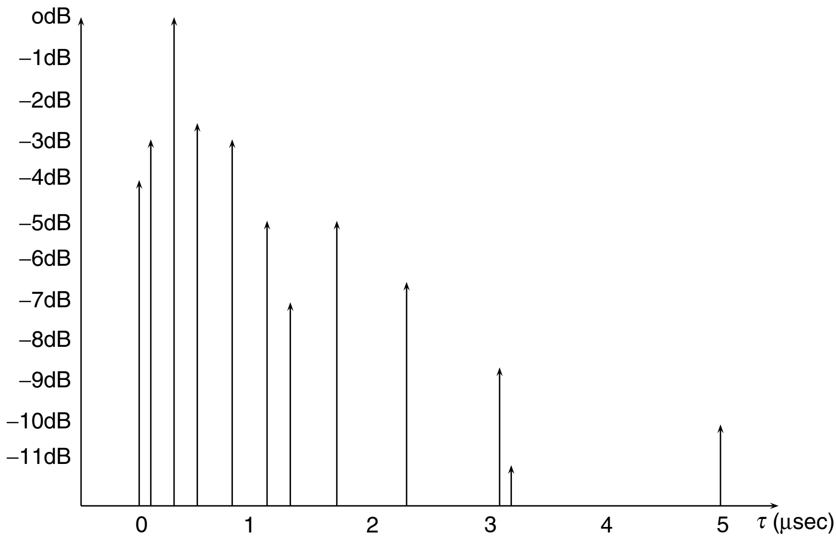


FIGURE 31.7 Power delay profile for the GSM model for typical urban channel.

TABLE 31.1 Parameter for Power Delay Profile GSM Model of Urban Area

Path	Delay (μs)	Average Power (dB)
1	0.0	−4.0
2	0.1	−3.0
3	0.3	0.0
4	0.5	−2.6
5	0.8	−3.0
6	1.1	−5.0
7	1.3	−7.0
8	1.7	−5.0
9	2.3	−6.5
10	3.1	−8.6
11	3.2	−11.0
12	5.0	−10.0

31.5 Propagation Loss

In this section we discuss the received power as a function of distance from the receiver. Suppose we have a transmitter and receiver separated by a distance d. The transmitter and receiver have antennas with gain G_t and G_r, respectively. If the transmitted power is P_t the received power is

$$P_r = P_t G_r G_t \left(\frac{\lambda}{4\pi d}\right)^2$$

where $\lambda = c/f$ is the wavelength of the signal. The above equation holds in free space without any reflections or multipath of any sort.

Now consider the case where there is an additional path due to a single reflection from the ground. The multipath has a different phase from the direct path. If we assume the reflection from the ground causes a

180° phase change then for large distances relative to the heights of the antennas the relation between the transmitted power and the received power changes to

$$P_r = P_t G_r G_t \frac{h_1^2 h_2^2}{d^4}$$

where h_1 and h_2 are the heights of the transmitting and receiving antenna. Thus, the relation of received power to distance becomes an inverse fourth power law or equivalently the power decreases 40 dB per decade of distance. Experimental evidence for wireless channels shows that the decrease in power with distance is 20 dB per decade near the base station but as the receiver moves away the rate of decrease increases. There are other models based on experimental measurements in different cities that give more complicated expressions for the path loss as a function of distance, antenna height, carrier frequency. See Reference 1 for further details.

31.6 Shadowing

In additon to the short-term multipath fading discussed earlier, long-term fading refers to shadowing of the receiver from the transmitter due to terrain and buildings. The time scale for long-term fading is much longer (on the order of seconds or minutes) than the time scale for short-term fading. Long-term fading is generally modeled as a log–normal random variable. In other words, the received power in dB has a normal (or Gaussian) distribution.

If a power measurement at a fixed distance from the transmitter was made there would be local variations due to constructive and destructive interference (short-term fading). In other words, what was discussed in Sections 31. 2–31. 4. At a fixed distance from the transmitter we would also have fluctuations in the received power because of the location of the receiver relative to various obstacles (e.g., buildings). If we were to measure the power over many locations separated by a distance of a wavelength or larger from a given point we would see that this average would vary depending on the location of measurement. Measurements with an obstacle blocking the direct line of sight path would have much smaller average than measurements without the obstacle. These fluctuations due to obstacles are called shadowing. Amplitude changes due to shadowing change much more slowly than amplitude changes that are due to multipath fading. Multipath fading causes amplitude and phase changes when the receiver moves about a wavelength (30 cm for a carrier frequency of 1 GHz) in distance while shadowing causes amplitude and phase fluctuations as the receiver moves about 10 m or more in distance.

The model for these shadowing-induced fluctuations is typically that of a log–normal distributed random variable for the received power. Equivalently, the power received expressed in dB is a Gaussian-distributed random variable with mean being the value determined by the propagation loss. The variance is dependent on the type of structures the mobile is located near and varies from about 3 to 6 dB. The fluctuations however are correlated. If $\nu(d)$ is a Gaussian random process modeling the shadowing process (in dB) at some location then the model for the correlation between the shadowing at distance d_1 and the shadowing at distance d_2 is

$$E\left[\nu(d_1)\nu(d_2)\right] = \sigma^2 \exp\{-|d_1 - d_2|/d_0\}$$

where d_0 is a parameter that determines how fast the correlation decays with distance. If the velocity is known then the correlation with time can be determined from the correlation in space. A typical value for d_0 is 10 m. Because shadowing is relatively slow it can be compensated for by power control algorithms.

31.7 Performance with (Time and Frequency) Nonselective Fading

In this section, we derive the performance of different modulation techniques with nonselective fading. We will ignore the propagation loss and shadowing effect and concentrate on the effects only due to multipath fading. First, the error probability conditioned on a particular fading level is determined. Then, the conditional error probability is averaged with respect to the distribution of the fading level.

31.7.1 Coherent Reception, Binary Phase Shift Keying

First, consider a modulator transmitting a binary phase shift keying (BPSK) signal and received with a faded amplitude. The transmitted signal is

$$s(t) = \sqrt{2P}b(t)\cos(2\pi f_c t)$$

where $b(t)$ is a data bit signal consisting of a sequence of rectangular pulses of amplitude $+1$ or -1. In other words,

$$b(t) = \sum_l b_l p_T(t - lT)$$

where $p_T(t)$ is a unit amplitude pulse of duration T beginning at time 0 and $b_l \in \{\pm 1\}$ are the data bits. The received signal is

$$r(t) = R\sqrt{2P}b(t)\cos(2\pi f_c t + \phi) + n(t)$$

where $n(t)$ is additive white Gaussian noise with two side power spectral density $N_0/2$. Assuming the receiver can accurately estimate the phase the demodulator (matched filter) output at time kT is

$$z_k = R\sqrt{E}b_{k-1} + \eta_k$$

where $E = PT$ is the transmitted energy, b_{k-1} is the data bit transmitted during the time interval $[(k-1)T, kT]$ and η_k is a Gaussian random variable with mean 0 and variance $N_0/2$. The random variable R represents the attenuation due to fading ($R = |X|$) or fading level and has probability density

$$p_R(r) = \begin{cases} 0, & r < 0 \\ \frac{r}{\sigma^2}\mathrm{e}^{-r^2}/2\sigma^2, & r \geq 0. \end{cases}$$

The density function determines the probability that the fading is between any two levels as

$$P\{a < R \leq b\} = \int_a^b p_R(r)\mathrm{d}r.$$

The error probability for a given fading level R is

$$P_\mathrm{e}(R) = Q\left(\sqrt{\frac{2ER^2}{N_0}}\right).$$

Here $Q(x) = \int_x^\infty \frac{1}{\sqrt{2\pi}}\mathrm{e}^{(-u^2/2)du}$

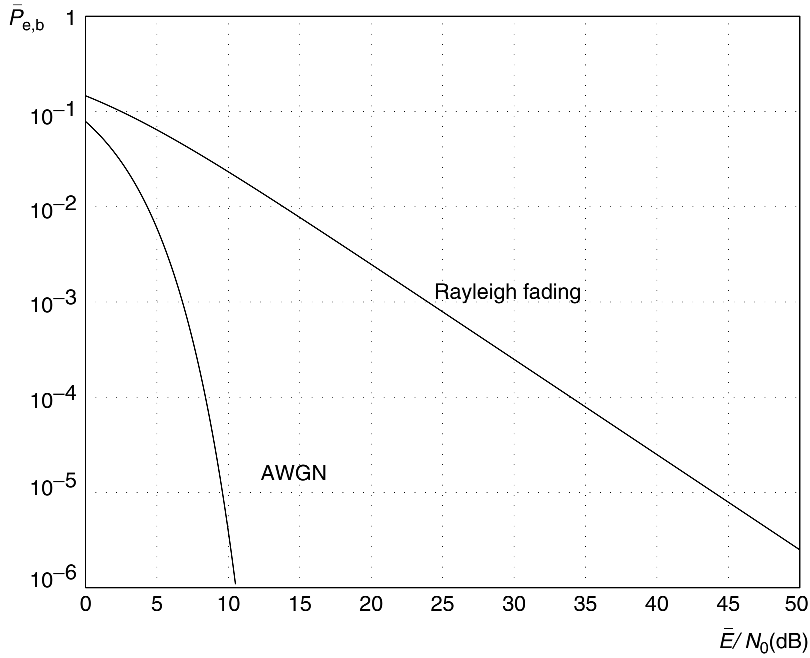


FIGURE 31.8 Bit error probability for BPSK with Rayleigh fading.

The unconditional error probability is the average of the conditional error probability for a given fade level with respect to the density of the fading level.

$$P_e = \int_{r=0}^{\infty} p_R(r) Q\left(\sqrt{\frac{2Er^2}{N_0}}\right) \mathrm{d}r \tag{31.1}$$

$$= \frac{1}{2} - \frac{1}{2}\sqrt{\frac{\bar{E}/N_0}{1 + E/N_0}}. \tag{31.2}$$

The error probability is shown in Figure 31.8 for the case of no fading (additive white Gaussian noise) and Rayleigh fading. For the additive white Gaussian noise channel the error probability decreases exponentially with signal-to-noise ratio, E/N_0. However, with fading the decrease in error probability is much slower. In fact, for large E/N_0 the error probability is

$$P_e \simeq \frac{1}{4E/N_0}.$$

Thus for high E/N_0 the error probability decreases inverse linearly with signal-to-noise ratio. To achieve an error probability of 10^{-5} requires a signal-to-noise ratio of 44.0 dB whereas in additive white Gaussian noise the required signal-to-noise ratio for the same error probability is 9.6 dB. Thus fading causes a loss in signal-to-noise ratio of 34.4 dB. This loss in performance is at the same average received power. The cause of this loss is the fact that the signal amplitude sometimes is very small and causes the error probability to be close to 1/2. Of course, sometimes the signal amplitude is large and results in very small error probability (say 0). However when we average the error probability, the result is much larger than the error probability at the average signal-to-noise ratio because of the highly nonlinear nature of the error probability as a function of signal amplitude without fading.

While the specific error probabilities change when the modulation changes the general nature of the error probabilities remain the same. In other words, without fading the error probability decreases exponentially with signal-to-noise ratio while with fading the error probability decreases inverse linearly with signal-to-noise ratio. This typically causes a loss in performance of between 30 and 40 dB and forces a designer into considering mitigation techniques as will be discussed subsequently.

31.7.2 BPSK with Diversity

To overcome this loss in performance (without just increasing power) a number of techniques can be applied. Many of the techniques attempt to receive the same information with independent fading statistics. This is generally called diversity. The diversity could be the form of L different antennas suitably separated so that the fading on different paths from the transmitter are independent. The diversity could be the form of transmitting the same data L times suitably separated in time so that the fading is independent.

In any case consider a system with L independent paths. The receiver demodulates each path coherently. Assume that the receiver also knows exactly the faded amplitude on each path. The decision statistics are then given by

$$z_l = r_l \sqrt{E} b + \eta_l, \quad l = 1, 2, \ldots, L$$

where r_l are Rayleigh, η_l are Gaussian and b represents the data bit transmitted which is either +1 or -1. The optimal method to combine the demodulator outputs can be derived as follows. Let $p_1(z_1, \ldots, z_L | r_1, \ldots, r_L)$ be the conditional density function of $z_1, \ldots, z_L$ given the transmitted bit is +1 and the fading amplitude is $r_1, \ldots, r_L$. The unconditional density is

$$p_1(z_1, \ldots, z_L, r_1, \ldots, r_L) = p_1(z_1, \ldots, z_L | r_1, \ldots, r_L) p(r_1, \ldots, r_L)$$

The conditional density of z_1 given $b = 1$ and r_1, is Gaussian with mean $r_l \sqrt{E}$ and variance $N_0/2$. The joint distribution of $z_1, \ldots, z_L$ is the product of the marginal density functions. The optimal combining rule is derived from the ratio

$$\begin{aligned} \Lambda &= \frac{p_1(z_1, \ldots, z_L, r_1, \ldots, r_L)}{p_{-1}(z_1, \ldots, z_L, r_1, \ldots, r_L)} \\ &= \exp\left\{ \frac{4}{N_0} \sum_{l=1}^{L} z_l r_l \sqrt{E} \right\}. \end{aligned}$$

The optimum decision rule is to compare Λ with 1 to make a decision. Thus the optimal rule is

$$\sum_{l=1}^{L} r_l Z_l \underset{b=-1}{\overset{b=+1}{\gtrless}} 0$$

This combining rule is known as maximum ratio combining. To implement this rule it is required that the receiver know the fading level for each transmission. The error probability with diversity L can be determined using the same technique as used without diversity. The expression for error probability is

$$P_e(L) = P_e(1) - \frac{1}{2} \sum_{k=1}^{L-1} \frac{(2k)!}{(k!k!)} (1 - 2P_e(1))(P_e(1))^k (1 - P_e(1))^k.$$

Here $P_e(1)$ is the error probability with diversity 1 which is given in Equation 31.2. The error probability as a function of the signal-to-noise ratio is shown in Figure 31.9. The signal-to-noise ratio in this case is

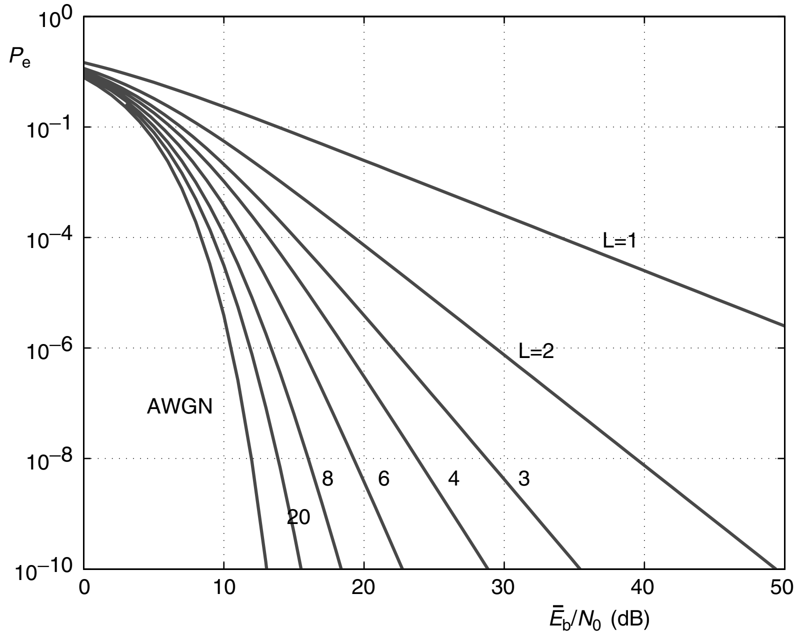


FIGURE 31.9 Error probability for BPSK (coherent demodulation) with and without Rayleigh fading.

defined as $E_b/N_0 = EL/N_0$ where E is the energy transmitted per transmitting antenna or time diversity. Thus E_b is the energy received per bit of information with L independent fading amplitudes. If we had L receiving antennas then the performance as a function of the transmitting energy would be L times better. In any case, we plot the error probability as a function of the total received energy. In the case of diversity transmission the energy transmitted per bit E_b is LE. For a fixed E_b as L increases each transmission contains less and less energy but there are more transmissions over independent faded paths. In the limit, as L becomes large (and using the weak law of large numbers) it can be shown that

$$\lim_{L\to\infty} P_e(L) = Q\left(\sqrt{\frac{2\bar{E}_b}{N_0}}\right)$$

For large signal-to-noise ratio the error probability with diversity L is decreases as $1/(E_b/N_0)^L$. While these curves show it is possible to get back to the performance with additive white Gaussian noise by using sufficient resources (diversity) it is possible to do even better with the right coding. In Figure 31.10, we show the performance (bit error probability) $P_{e,b}$ of a rate 1/2 constraint length 7 convolutional code on a Rayleigh faded channel (independent fading on each bit) where the receiver knows the fading level (side information) for each bit and can appropriately weight the metric in the decoder. Notice that the required E_b/N_0 for 10^{-5} bit error probability is about 7.5 dB, which is less than that required for uncoded BPSK without fading. The gain compared to uncoded performance is more than 36 dB.

31.7.3 Fundamental Limits

Diversity order L, as described above, requires that the data be repeated L times. This decreases the data rate by a factor of L for a given bandwidth. Other approaches to achieving diversity are possible and include using L receiving antennas or using some combination of transmitting and receiving antennas. This is generally known as multiple-input multiple-output (MIMO). For a single antenna it is of interest to know the maximum data rate for which arbitrarily small error probability can be achieved. The

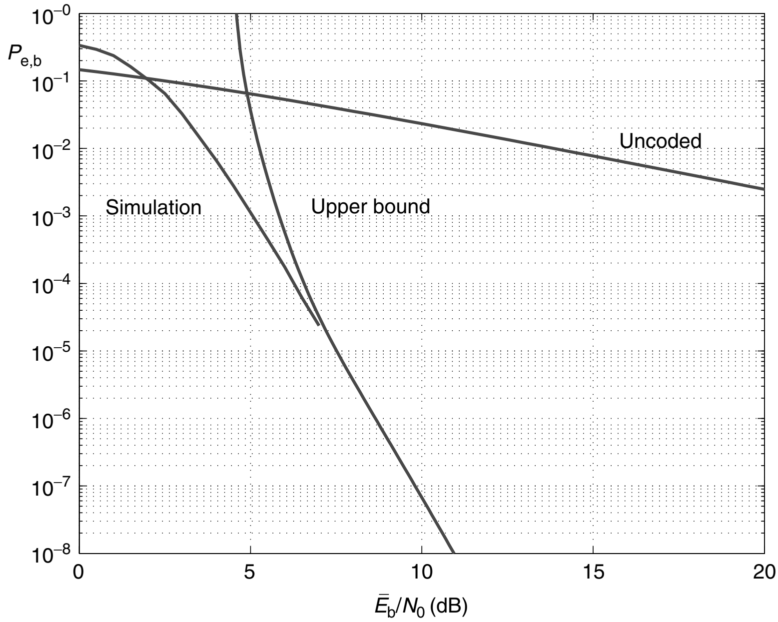


FIGURE 31.10 Error probability for BPSK (coherent demodulation) with Rayleigh fading and convolutional coding.

fundamental limit on performance can be determined for a variety of circumstances. Here we assume that the transmitter has no knowledge of the fading amplitude and that the modulation is BPSK. When the receiver knows exactly what the amplitude (and phase) of the fading process is, we say that, side information is available. The maximum rate of transmission (in bits/symbol) is called the capacity of the channel C. If an error control code of rate r information bits/channel use is used then reliable (arbitrarily small error probability communication) is possible provide the rate is less than the capacity. For the case of side information available this condition is

$$r < C = 1 - \int_{r=0}^{\infty} \int_{y=-\infty}^{\infty} f(r) g(y) \log_2(1 + \mathrm{e}^{-2y\beta}) \mathrm{d}y \mathrm{d}r$$

where $f(r) = 2r \exp\{-r^2\}$, $\beta = \sqrt{2E/N_0}$ and

$$g(y) = \frac{1}{\sqrt{2\pi}} \exp\left\{ - \left(y - \sqrt{2\bar{E}r^2/N_0} \right)^2 \Big/ 2 \right\}.$$

If the receiver does not know the fading amplitude (but still does coherent demodulation) then we say no side information is available. The rate at which reliable communication is possible in this case satisfies

$$r < C = 1 - \int_{r=0}^{\infty} \int_{y=-\infty}^{\infty} p(y|1) \log_2 \left(1 + \frac{p(y|0)}{p(y|1)} \right) \mathrm{d}y$$

where

$$p(y|0) = \int_0^{\infty} f(r) \frac{1}{\sqrt{2\pi N_0}} \mathrm{e}^{-(y-\sqrt{E}r)^2/N_0} \mathrm{d}r$$

and

$$p(y|1) = \int_0^\infty f(r)\frac{1}{\sqrt{2\pi N_0}}\mathrm{e}^{-(y+\sqrt{\bar{E}}r)^2/N_0}\mathrm{d}r$$

If the receiver makes a hard decision about each modulated symbol and the receiver knows the fading amplitude then the capacity is

$$C = \int_0^\infty f(r)[1 + p(r)\log_2(p(r)) + (1 - p(r))\log_2(1 - p))]\mathrm{d}r$$

where $p(r) = Q(\sqrt{2\bar{E}r^2/N_0})$. For a receiver that does not know the fading amplitude and makes hard decisions on each coded bit the capacity is given by

$$C = 1 + \bar{p}\log_2(\bar{p}) + (1 - \bar{p})\log_2(1 - \bar{p})$$

where

$$\bar{p} = \frac{1}{2} - \frac{1}{2}\frac{\sqrt{\bar{E}/N_0}}{E/N_0 + 1}.$$

Finally if the transmitter is not restricted to BPSK but can use any type of modulation then the capacity when the receiver knows the fading level is

$$C = \int_0^\infty f(r)\frac{1}{2}\log_2(1 + 2\bar{E}r^2/N_0)\mathrm{d}r$$

Because the capacity depends on the amount of energy used per transmitted bit we can reinterpret the capacity result in the following way. For a given code rate there is a minimum energy required for the capacity to be larger than the code rate. Usually the energy per information bit E_b is more important than the energy per transmitted bit. Since for a code of rate r, there are r information bits per coded bit we can relate the energy per transmitted (coded) bit and information bit as

$$E_b = \frac{E}{r}.$$

Using this we can interpret the capacity result as a minimum energy per information bit.

$$\begin{aligned} C(E/N_0) &> r, \\ E/N_0 &> C^{-1}(r), \\ E_b/N_0 &> C^{-1}(r)/r. \end{aligned}$$

In Figure 31.11, we show the minimum signal-to-noise ratio E_b/N_0 per information bit required for arbitrarily reliable communication as a function of the code rate r being used. In this figure the top curve (a) is the minimum signal-to-noise ratio necessary for reliable communication with hard decisions and no side information. The second curve (b) is the case of hard decisions with side information. The third curve (c) is the case of soft decisions with side information and binary modulation (BPSK). The bottom curve (d) is the case of unrestricted modulation and side information available at the receiver. There is about a 2 dB gap between hard decision and soft decisions when side information is available. There is an extra 1 dB degradation in hard decisions if the receiver does not know the amplitude. A roughly similar degradation in performance is also true for soft decisions with and without side information. The model

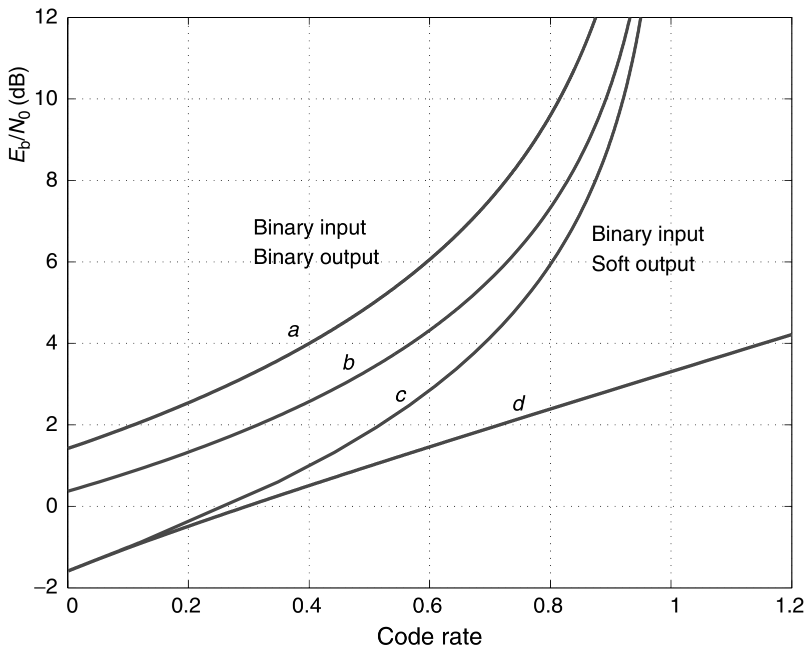


FIGURE 31.11 Capacity of Rayleigh faded channel with coherent detection.

shown here assumes that the fading is constant over one symbol duration, but independent from one symbol to the next. However, for the case of the receiver knowing the fading level (side information) the capacity actually does not depend on the time selectivity as long as the fading is constant for at least one symbol duration. When there is no side information, the capacity gets larger when there is less selectivity. In this case the receiver can better estimate the channel characteristics and use the channel characteristics in demodulating the signal. In fact, as the channel coherence time becomes large, the capacity *without* side information approaches the capacity *with* side information.

As can be seen from Figure 31.11 it is extremely important that some form of coding be used with fading. The required signal-to-noise ratio for small error probabilities has decreased from on the order of 45 dB for an uncoded systems to a mere 2 dB for a coded system. When a repetition code is used the error probability can be made to decrease exponentially with signal to-noise ratio provided that we use a large number of antennas or repeat the same symbol a large number of times which results in a small rate of transmission (information bits/modulated symbol). However, with error control coding (such as a convolutional code) and independent fading we can greatly improve the performance without a significant decrease in the rate of transmission. The minimum required signal-to-noise ratio is no different from an unfaded channel when very low rate coding is used. For rate 1/2 coding, the loss in performance is less than 2 dB compared to an unfaded channel.

In conclusion, multipath fading causes the signal amplitude to vary and the performance of typical modulation techniques to degrade by tens of dB. However, with the right amount of error control coding the required signal-to-noise ratio can be decreased to within 2 dB of the required signal-to-noise ratio for an additive white Gaussian channel when the code rate is 0.5.

Reference

1. Kaveh Pahlavan and Allen H. Levesque, *Wireless Information Networks*, Wiley, 1995.

32

Electromagnetic Interference (EMI)

Alfy Riddle
Finesse, LLC

32.1 Fundamentals of EMI

Electromagnetic interference is a potential hazard to all wireless and wired products. Most EMI concerns are due to one piece of equipment unintentionally affecting another piece of equipment, but EMI problems can arise within an instrument as well. Often the term electromagnetic compatibility (EMC) is used to denote the study of EMI effects. The following sections on generation of EMI, shielding of EMI, and probing for EMI will be helpful in both internal product EMI reduction and external product EMI compliance.

EMI compliance is regulated in the United States through the Federal Communications Commission (FCC). Specifically, Parts 15 and 18 of the Code of Federal Regulations (CFR) govern radiation standards and standards for industrial, scientific, and medical equipment. In Europe, Publication 22 from the Comite International Special des Perturbations Radioelectriques (CISPR) governs equipment radiation. Although the primary concern is compliance with radiation standards, conduction of unwanted signals onto power lines causes radiation from the long power lines, so conducted EMI specifications are also included in FCC Part 15 and CISPR 22 [1,2].

Figure 32.1 shows the allowed conducted EMI. FCC and CISPR specifications do not set any limits above 30 MHz. All of the conducted measurements are to be done with a line impedance stabilization network (LISN) connected in the line. The LISN converts current-based EMI to a measurable voltage. The LISN uses series inductors of 50 μH to build up a voltage from line current interference, and 0.1 μF capacitors couple the noise voltage to 50 Ω resistors for measurement [1]. Capacitors of 1 μF also bridge the output so the inductors see an AC short. The measurements in Figure 32.1 are reported in dBμV, which is dB with respect to 1 μV. Both FCC and CISPR measurements specify an RF bandwidth of at least 100 kHz. The CISPR limitations given in Figure 32.1 denote that a quasi-peak (QP) detector should be used. The QP detector is more indicative of human responses to interference. CISPR specifications for an averaging detector are 10 dB below that of the QP detector. FCC specifications require a QP detector.

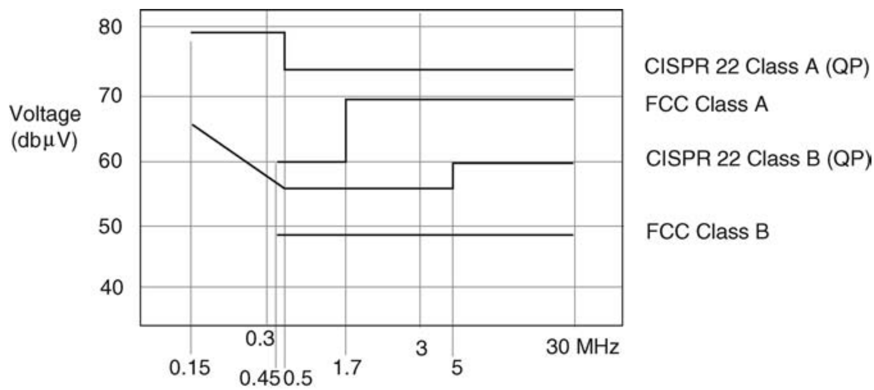


FIGURE 32.1 Conducted EMI specifications, measured with LISN.

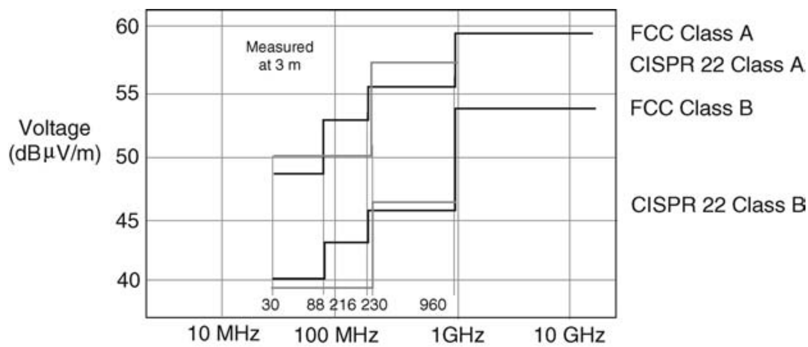


FIGURE 32.2 Radiated EMI specifications referred to 3 m.

Both FCC and CISPR limitations have two classes. Class A is basically for industrial use, while Class B is for residential use.

Both CISPR and FCC radiated EMI specifications begin at 30 MHz. Figure 32.2 shows the radiation limits for CISPR and FCC Classes A and B [1,2]. Because these measurements are made with an antenna, they are specified as a field strength in dB referenced to 1 µV/m. The measurement distances for radiation limits varies in the specifications, but all of the limits shown in Figure 32.2 are referred to 3 m. Other distances can be derived by reducing the limits by 20 dB for every factor of ten increase in distance.

32.2 Generation of EMI

Almost any component can generate EMI. Oscillators, digital switching circuits, switching regulators, and fiber-optic transmitters can radiate through PCB traces, inductors, gaps in metal boxes, ground loops, and gaps in ground planes [1].

32.2.1 Switching Regulators

Because switching regulators have a fundamental frequency below 30 MHz and switch large currents at high speed, they can contribute to both conducted and radiated emissions. Very careful filtering is required so switching regulators do not contaminate ground planes with noise. The input filters on switching regulators are just as important as the output filters because the noise can travel to other power supplies and out to the line [3].

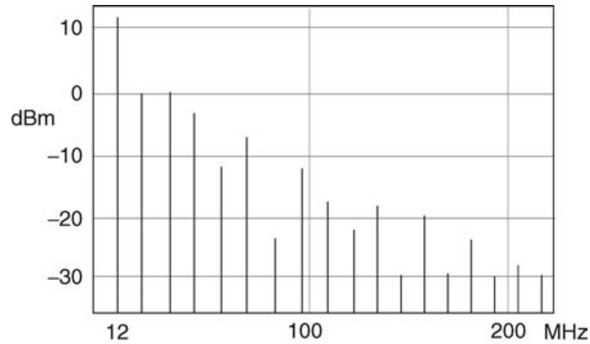


FIGURE 32.3 Harmonic clock spectra.

32.2.2 Digital Switching

Digital networks have several characteristics that increase EMI. Digital networks tend to have many lines switching simultaneously. The currents on the lines add in phase and increase radiation. Also, as CMOS digital circuits increase in clock frequency they require more current to drive their loads. Increasing both the current and the frequency creates more di/dt noise through the inductive connection from ICs to ground [4]. The fast switching of digital waveforms produces harmonics decades beyond the oscillator fundamental frequency. Figure 32.3 shows the spectra from a typical 12 MHz crystal square-wave clock oscillator. Note that even though the output appears as a square wave, both even and odd harmonics are present. The harmonics of this oscillator fall off at roughly 20 dB/decade. As will be seen in the next section, most sources of coupling increase at about 20 dB/decade, which causes not only a relatively flat coupling spectrum, but significant EMI at ten or even 100 times the oscillator frequency. While this 12 MHz clock oscillator is at a very low frequency, similar phenomena happens with the laser drivers for high speed fiber-optic networks that operate with clocks of 2.5 GHz and higher.

32.3 Coupling

Any inductor is a potential source of coupling and radiation. Ribbon cables act like very long coupling loops and can spread signals or power supply noise all over an instrument and out to the outside world. Fundamentally, lengths of wire radiate an electric field and loops of wire radiate a magnetic field. The electric field, E_{Far}, far from a short radiator, is given by Equation 32.1 [1]. Equation 32.1 is in volts per meter where I is the element current, l is the element length, λ is the radiation wavelength, and r is the distance from the radiating element to the measured field. Because the field strength is inversely proportional to frequency, EMI coupling tends to increase with frequency. The far magnetic field, H_{Far}, due to a current loop is given in A/m by Equation 32.2. In Equation 32.2 a is the radius of the loop, c is the speed of light, I is the current in the loop, and r is the distance from the loop to the measured radiation.

$$\mathrm{E}_{\mathrm{Far}} = 377\ \mathrm{I}l/(2\lambda r) \tag{32.1}$$

$$\mathrm{H}_{\mathrm{Far}} = \omega^2 a^2 \mu \mathrm{I}/(1508 c r) \tag{32.2}$$

32.3.1 Cabling

Cables form a source of radiation and susceptibility. In general it is the signal on the shield of a coaxial cable or the common-mode signal on twisted pairs that generates most of the radiation [1]. However, even high-quality coaxial cables will leak a finite amount of signal through their shields. Multiple braids, solid outer conductors, and even solder-filled braid coaxial cables are used to increase shielding. Twisted pair cables rely on twisting to cancel far field radiation from the differential mode. The effectiveness of the twisting reduces as frequency increases [1].

32.4 Shielding

Shielding is the basic tool for EMC work. Shielding keeps unwanted signals out and potential EMI sources contained. For the most part, a heavy metal box with no seams or apertures is the most effective shield. While thin aluminum enclosures with copper tape over the seams appear to enclose the RF currents, in fact they are a poor substitute for a heavy gauge cast box with an EMI gasket. It has been said that if spectrum analyzer manufacturers could make their instruments any lighter, primarily by leaving out some of the expensive casting, they would. The basic equation for shielding is given by [5].

$$S = A + R + B \tag{32.3}$$

In Equation 32.3, A is the shield absorption in dB, R is the shield reflection in dB, and B is a correction factor for multiple reflections within the shield [5]. Shield effectiveness depends on the nature of the field. Purely electric fields are well isolated by thin conductive layers while purely magnetic fields are barely attenuated by such layers. Magnetic fields require thick layers of high permeability material for effective shielding at low frequencies. Plane waves contain a fairly high impedance mix of electric and magnetic fields that are both reflected and absorbed by thin metal layers provided the frequency is high enough. One of the subtle points in EMI shielding is that any slot can destroy shielding effectiveness. It is the length of a slot in comparison to a wavelength that determines how easily a wave can pass through the slot [5].

32.5 Measurement of EMI

EMI compliance must be verified. Unfortunately, EMI measurements are time consuming, tedious, plagued by local interference sources, and often of frustrating variability. The FCC requires measurements to be verified at an Open Area Test Site (OATS) [2]. Many manufactures use a local site or a shielded TEM cell to estimate FCC compliance during product development. With care, OATS and TEM cell measurements can be correlated [6]. In any case, even careful EMI measurements can wander by several dB so most manufacturers design for a healthy margin in their products.

32.5.1 Open Area Test Site (OATS)

A sketch of an OATS is given in Figure 32.4a. OATS testing involves setting the antenna at specified distances from the device under test (DUT). FCC and CISPR regulations use distances of 3, 10, and 30 m

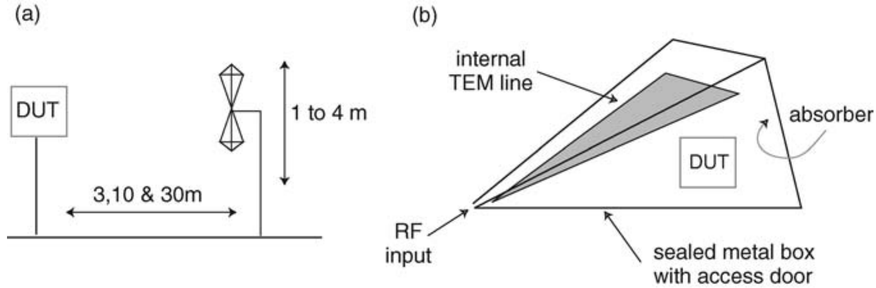


FIGURE 32.4 EMI measurements methods: (a) OATS; and (b) TEM cell.

depending on the verification class [1,2]. The antenna height must also be varied to account for ground reflections. Finally, the DUT must be rotated about all axes and the antenna must be utilized to test the DUT under radiation by both horizontally and vertically polarized fields.

32.5.2 TEM Cell

TEM cells are a convenient and relatively low cost method for making accurate and well-isolated EMI measurements [2,7]. A sketch of one configuration of TEM cell is shown in Figure 32.4b [2]. The TEM cell uses a transmission line in a box to create a TEM field for testing devices. The box is driven from a narrow end and usually expands into an area where the DUT can be placed. The box terminates in a resistor surrounded by RF absorber. For EMI measurements the DUT must be placed away from the box walls and rotated about each axis so that all possible radiated waves can be measured.

32.5.3 Probes

EMI probes can be a very effective way of solving EMI problems. Articles have been written on building probes and commercial probes are available that provide a flat frequency response [8]. These probes can be used to "sniff" around a device until signals with the same spectral spacing as the EMI problem can be found. For clocks and switching power supplies, a close examination of the spectral spacing will indicate the fundamental frequency, which can be traced to a component on a schematic. In the time domain the suspected EMI source can be used to trigger an oscilloscope with the probed signal as the oscilloscope input. If the trace is stable, then the suspected EMI source has been found. Electric field probes are based on Equation 32.1 and can be as simple as a wire extending from a connector. Magnetic field probes can be as simple as a loop of wire completing the path from a connector's center pin to its flange. A rectangular loop of length l with the near side a distance a from a current source I, and having the far side a distance b from the current source will yield the voltage given in Equation 32.4. With all small probes it is useful to have at least a 6 dB pad after the probe to establish a load and minimize reflections.

$$V = j\omega l \mu I \ln(b/a)/(2\pi) \tag{32.4}$$

32.6 Summary

EMI problems create an inexhaustible supply of work for those in the field. The EMC field has well documented requirements and a long history of measurement. Many excellent sources of information exist for those working in this area [1,2,5,9].

References

1. Paul, C.R., *Introduction to Electromagnetic Compatibility*, John Wiley & Sons, New York, 1992.
2. Morgan, D., *A Handbook for EMC Testing and Measurement*, Peter Peregrinus Ltd., London, 1994.
3. Lee, F.C. and Yu, Y., Input-Filter Design for Switching Regulators, *IEEE Trans. Aerosp. and Electron. Sys.*, 627–634, September 1979.
4. Dolle, M., Analysis of Simultaneous Switching Noise, *IEEE ISCAS*, 904–907, 1995.
5. Ott, H.W., *Noise Reduction Techniques in Electronic Systems*, John Wiley & Sons, New York, 1976.
6. Wilson, P., On Correlating TEM Cell and OATS Emission Measurements, *IEEE Trans. EMC*, 1–16, February 1995.
7. Konigstein, D., and Hansen, D., A New Family of TEM-Cells with Enlarged bandwidth and Optimized Working Volume, *Proc. 7th Int'l. Zurich Symp. on EMC*, 127–132, March 1987.
8. Johnson, F., Simple "Homemade" Sensors Solve Tough EMI Problems, *Electron. Design*, 109–114, November 8, 1999.
9. Kodali, V.P. and Kanda, M., *EMC/EMI Selected Readings*, IEEE Press, Piscataway, NJ, 1996.

Appendix A
Mathematics, Symbols, and Physical Constants

Greek Alphabet

	Greek Letter	Greek Name	English Equivalent		Greek Letter	Greek Name	English Equivalent
Α	α	Alpha	a	Ν	ν	Nu	n
Β	β	Beta	b	Ξ	ξ	Xi	x
Γ	γ	Gamma	g	Ο	ο	Omicron	ŏ
Δ	δ	Delta	d	Π	π	Pi	p
Ε	ε	Epsilon	ĕ	Ρ	ρ	Rho	r
Ζ	ζ	Zeta	z	Σ	σ	Sigma	s
Η	η	Eta	ē	Τ	τ	Tau	t
Θ	θ ϑ	Theta	th	Υ	υ	Upsilon	u
Ι	ι	Iota	i	Φ	ϕ φ	Phi	ph
Κ	κ	Kappa	k	Χ	χ	Chi	ch
Λ	λ	Lambda	l	Ψ	ψ	Psi	ps
Μ	μ	Mu	m	Ω	ω	Omega	ō

International System of Units (SI)

The International System of units (SI) was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960. It is a coherent system of units built form seven *SI base units,* one for each of the seven dimensionally independent base quantities: they are the meter, kilogram, second, ampere, kelvin, mole, and candela, for the dimensions length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity, respectively. The definitions of the SI base units are given below. The *SI derived units* are expressed as products of powers of the base units, analogous to the corresponding relations between physical quantities but with numerical factors equal to unity.

In the International System there is only one SI unit for each physical quantity. This is either the appropriate SI base unit itself or the appropriate SI derived unit. However, any of the approved decimal prefixes, called *SI prefixes,* may be used to construct decimal multiples or submultiples of SI units.

It is recommended that only SI units be used in science and technology (with SI prefixes where appropriate). Where there are special reasons for making an exception to this rule, it is recommended always to define the units used in terms of SI units. This section is based on information supplied by IUPAC.

Definitions of SI Base Units

Meter—The meter is the length of path traveled by light in vacuum during a time interval of 1/299 792 458 of a second (17th CGPM, 1983).

Kilogram—The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).

Second—The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967).

Ampere—The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross-section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length (9th CGPM, 1948).

Kelvin—The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (13th CGPM, 1967).

Mole—The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, or other particles, or specified groups of such particles (14th CGPM, 1971).

Examples of the use of the mole:

1 mol of H_2 contains about 6.022×10^{23} H_2 molecules, or 12.044×10^{23} H atoms
1 mol of HgCl has a mass of 236.04 g
1 mol of Hg_2Cl_2 has a mass of 472.08 g
1 mol of Hg_2^{2+} has a mass of 401.18 g and a charge of 192.97 kC
1 mol of $Fe_{0.91}S$ has a mass of 82.88 g
1 mol of e^- has a mass of 548.60 µg and a charge of –96.49 kC
1 mol of photons whose frequency is 10^{14} Hz has energy of about 39.90 kJ

Candela—The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} hertz and that has a radiant intensity in that direction of (1/683) watt per steradian (16th CGPM, 1979).

Names and Symbols for the SI Base Units

Physical Quantity	Name of SI Unit	Symbol for SI Unit
Length	Meter	m
Mass	Kilogram	kg
Time	Second	s
Electric Current	Ampere	A
Thermodynamic temperature	Kelvin	K
Amount of substance	Mole	mol
Luminous intensity	Candela	cd

SI Derived Units with Special Names and Symbols

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units	
Frequency[1]	Hertz	Hz	s^{-1}	
Force	Newton	N	$m\ kg\ s^{-2}$	
Pressure, stress	Pascal	Pa	$N\ m^{-2}$	$= m^{-1}\ kg\ s^{-2}$
Energy, work, heat	Joule	J	N m	$= m^2\ kg\ s^{-2}$
Power, radiant flux	Watt	W	$J\ s^{-1}$	$= m^2\ kg\ s^{-3}$
Electric charge	Coulomb	C	A s	
Electric potential, electromotive force	Volt	V	$J\ C^{-1}$	$= m^2\ kg\ s^{-3}\ A^{-1}$

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units	
Electric resistance	Ohm	Ω	$V\ A^{-1}$	$= m^2\ kg\ s^{-3}\ A^{-2}$
Electric conductance	Siemens	S	Ω^{-1}	$= m^{-2}\ kg^{-1}\ s^3\ A^2$
Electric capacitance	Farad	F	$C\ V^{-1}$	$= m^{-2}\ kg^{-1}\ s^4\ A^2$
Magnetic flux density	Tesla	T	$V\ s\ m^{-2}$	$= kg\ s^{-2}\ A^{-1}$
Magnetic flux	Weber	Wb	V s	$= m^2\ kg\ s^{-2}\ A^{-1}$
Inductance	Henry	H	$V\ A^{-1}\ s$	$= m^2\ kg\ s^{-2}\ A^{-2}$
Celsius temperature[2]	Degree Celsius	°C	K	
Luminous flux	Lumen	lm	cd sr	
Illuminance	Lux	lx	$cd\ sr\ m^{-2}$	
Activity (Radioactive)	Becquerel	Bq	s^{-1}	
Absorbed dose (of radiation)	Gray	Gy	$J\ kg^{-1}$	$= m^2\ s^{-2}$
Dose equivalent (dose equivalent index)	Sievert	Sv	$J\ kg^{-1}$	$= m^2\ s^{-2}$
Plane angle	Radian	rad	1	$= m\ m^{-1}$
Solid angle	Steradian	sr	1	$= m^2\ m^{-2}$

[1]For radial (circular) frequency and for angular velocity the unit rad s^{-1}, or simply s^{-1}, should be used, and this may not be simplified to Hz. The unit Hz should be used only for frequency in the sense of cycles per second.

[2]The Celsius temperature θ is defined by the equation:

$$\theta/°C = T/K - 273.15$$

The SI unit of Celsius temperature interval is the degree Celsius, °C, which is equal to the kelvin, K. °C should be treated as a single symbol, with no space between the ° sign and the letter C. (The symbol °K, and the symbol °, should no longer be used.)

Units in Use Together with the SI

These units are not part of the SI, but it is recognized that they will continue to be used in appropriate contexts. SI prefixes may be attached to some of these units, such as milliliter, ml; millibar, mbar; megaelectronvolt, MeV; kilotonne, ktonne.

Physical Quantity	Name of Unit	Symbol for Unit	Value in SI Units
Time	Minute	min	60 s
Time	Hour	h	3600 s
Time	Day	d	86 400 s
Plane angle	Degree	°	$(\pi/180)$ rad
Plane angle	Minute	′	$(\pi/10\ 800)$ rad
Plane angle	Second	″	$(\pi/648\ 000)$ rad
Length	Ångstrom[1]	Å	10^{-10} m
Area	Barn	b	$10^{-28}\ m^2$
Volume	Litre	l, L	$dm^3 = 10^{-3}\ m^3$
Mass	Tonne	t	$Mg = 10^3$ kg
Pressure	Bar[1]	bar	$10^5\ Pa = 10^5\ N\ m^{-2}$
Energy	Electronvolt[2]	eV (= $e \times$ V)	$\approx 1.60218 \times 10^{-19}$ J
Mass	Unified atomic mass unit[2,3]	u (= $m_a(^{12}C)/12$)	$\approx 1.66054 \times 10^{-27}$ kg

[1]The ångstrom and the bar are approved by CIPM for "temporary use with SI units," until CIPM makes a further recommendation. However, they should not be introduced where they are not used at present.

[2]The values of these units in terms of the corresponding SI units are not exact, since they depend on the values of the physical constants e (for the electronvolt) and N_a (for the unified atomic mass unit), which are determined by experiment.

[3]The unified atomic mass unit is also sometimes called the dalton, with symbol Da, although the name and symbol have not been approved by CGPM.

Physical Constants

General

Equatorial radius of the earth = 6378.388 km = 3963.34 miles (statute).
Polar radius of the earth, 6356.912 km = 3949.99 miles (statute).
1 degree of latitude at 40° = 69 miles.
1 international nautical mile = 1.15078 miles (statute) = 1852 m = 6076.115 ft.
Mean density of the earth = 5.522 g/cm^3 = 344.7 lb/ft^3.
Constant of gravitation $(6.673 \pm 0.003) \times 10^{-8}$ cm^3 gm^{-1} s^{-2}.
Acceleration due to gravity at sea level, latitude 45° = 980.6194 cm/s^2 = 32.1726 ft/s^2.
Length of seconds pendulum at sea level, latitude 45° = 99.3575 cm = 39.1171 in.
1 knot (international) = 101.269 ft/min = 1.6878 ft/s = 1.1508 miles (statute)/h.
1 micron = 10^{-4} cm.
1 ångstrom = 10^{-8} cm.
Mass of hydrogen atom = $(1.67339 \pm 0.0031) \times 10^{-24}$ g.
Density of mercury at 0°C = 13.5955 g/ml.
Density of water at 3.98°C = 1.000000 g/ml.
Density, maximum, of water, at 3.98°C = 0.999973 g/cm^3.
Density of dry air at 0°C, 760 mm = 1.2929 g/l.
Velocity of sound in dry air at 0°C = 331.36 m/s – 1087.1 ft/s.
Velocity of light in vacuum = $(2.997925 \pm 0.000002) \times 10^{10}$ cm/s.
Heat of fusion of water 0°C = 79.71 cal/g.
Heat of vaporization of water 100°C = 539.55 cal/g.
Electrochemical equivalent of silver 0.001118 g/s international amp.
Absolute wavelength of red cadmium light in air at 15°C, 760 mm pressure = 6438.4696 Å.
Wavelength of orange-red line of krypton 86 = 6057.802 Å.

π Constants

π = 3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37511
$1/\pi$ = 0.31830 98861 83790 67153 77675 26745 02872 40689 19291 48091
π^2 = 9.8690 44010 89358 61883 44909 99876 15113 53136 99407 24079
$\log_e \pi$ = 1.14472 98858 49400 17414 34273 51353 05871 16472 94812 91531
$\log_{10} \pi$ = 0.49714 98726 94133 85435 12682 88290 89887 36516 78324 38044
$\log_{10} \sqrt{2}\,\pi$ = 0.39908 99341 79057 52478 25035 91507 69595 02099 34102 92128

Constants Involving *e*

e = 2.71828 18284 59045 23536 02874 71352 66249 77572 47093 69996
$1/e$ = 0.36787 94411 71442 32159 55237 70161 46086 74458 11131 03177
e^2 = 7.38905 60989 30650 22723 04274 60575 00781 31803 15570 55185
$M = \log_{10} e$ = 0.43429 44819 03251 82765 11289 18916 60508 22943 97005 80367
$1/M = \log_e 10$ = 2.30258 50929 94045 68401 79914 54684 36420 67011 01488 62877
$\log_{10} M$ = 9.63778 43113 00536 78912 29674 98645 –10

Numerical Constants

$\sqrt{2}$ = 1.41421 35623 73095 04880 16887 24209 69807 85696 71875 37695
$3\sqrt{2}$ = 1.25992 10498 94873 16476 72106 07278 22835 05702 51464 70151
$\log_e 2$ = 0.69314 71805 59945 30941 72321 21458 17656 80755 00134 36026
$\log_{10} 2$ = 0.30102 99956 63981 19521 37388 94724 49302 67881 89881 46211
$\sqrt{3}$ = 1.73205 08075 68877 29352 74463 41505 87236 69428 05253 81039
$\sqrt[3]{3}$ = 1.44224 95703 07408 38232 16383 10780 10958 83918 69253 49935
$\log_e 3$ = 1.09861 22886 68109 69139 52452 36922 52570 46474 90557 82275
$\log_{10} 3$ = 0.47712 12547 19662 43729 50279 03255 11530 92001 28864 19070

Symbols and Terminology for Physical and Chemical Quantities

Name	Symbol	Definition	SI Unit
		Classical Mechanics	
Mass	m		kg
Reduced mass	μ	$\mu = m_1 m_2/(m_1 + m_2)$	kg
Density, mass density	ρ	$\rho = M/V$	kg m^{-3}
Relative density	d	$d = \rho/\rho^\theta$	l
Surface density	ρ_A, ρ_S	$\rho_A = m/A$	kg m^{-2}
Specific volume	ν	$\nu = V/m = 1/\rho$	m^3 kg^{-1}
Momentum	p	$\boldsymbol{p} = mv$	kg m s^{-1}
Angular momentum, action	L	$\mathrm{l} = \mathrm{r} \times \mathrm{p}$	J s
Moment of inertia	I, J	$I = \Sigma m_i r_i^2$	kg m^2
Force	F	$\boldsymbol{F} = \mathrm{d}\boldsymbol{p}/\mathrm{d}t = ma$	N
Torque, moment of a force	$\boldsymbol{T}$, ($\boldsymbol{M}$)	$\mathrm{T} = \mathrm{r} \times \mathrm{F}$	N m
Energy	E		J
Potential energy	E_p, V, Φ	$E_\mathrm{p} = -\int \boldsymbol{F} \cdot \mathrm{d}s$	J
Kinetic energy	E_k, T, K	$e_\mathrm{k} = (1/2)mv^2$	J
Work	W, w	$w = \int \boldsymbol{F} \cdot \mathrm{d}s$	J
Hamilton function	H	$H(q, p)$ $= T(q, p) + V(q)$	J
Lagrange function	L	$L(q, \dot{q})$ $\mathrm{T}(q, \dot{q}) - V(q)$	J
Pressure	p, P	$p = F/A$	Pa, N m^{-2}
Surface tension	γ, σ	$\gamma = \mathrm{d}W/\mathrm{d}A$	N m^{-1}, J m^{-2}
Weight	G, (W, P)	$G = mg$	N
Gravitational constant	G	$F= Gm_1m_2/r^2$	N m^2 kg^{-2}
Normal stress	σ	$\sigma = F/A$	Pa
Shear stress	τ	$\tau = F/A$	Pa
Linear strain, relative elongation	ε, e	$\varepsilon = \Delta l/l$	l
Modulus of elasticity, Young's modulus	E	$E = \sigma/\varepsilon$	Pa
Shear strain	γ	$\gamma = \Delta x/d$	l
Shear modulus	G	$\mathrm{G} = \tau/\gamma$	Pa
Volume strain, bulk strain	θ	$\theta = \Delta V/V_0$	l
Bulk modulus, compression modulus	K	$K = -V_0(\mathrm{d}p/\mathrm{d}V)$	Pa
Viscosity, dynamic viscosity	η, μ	$\tau_{x,z} = \eta(\mathrm{d}v_x/\mathrm{d}z)$	Pa s
Fluidity	ϕ	$\phi = 1/\eta$	m kg^{-1} s
Kinematic viscosity	ν	$\nu = \eta/\rho$	m^2 s^{-1}
Friction coefficient	μ, (f)	$F_\mathrm{frict} = \mu F_\mathrm{norm}$	l
Power	P	$P = \mathrm{d}W/\mathrm{d}t$	W
Sound energy flux	P, P_a	$P = \mathrm{d}E/\mathrm{d}t$	W
Acoustic factors			
Reflection factor	ρ	$\rho = P_\mathrm{t}/P_0$	l
Acoustic absorption factor	α_a, (α)	$\alpha_\mathrm{a} = 1 - \rho$	l
Transmission factor	τ	$\tau = P_\mathrm{tr}/P_0$	l
Dissipation factor	δ	$\delta = \alpha_\mathrm{a} - \tau$	l

Fundamental Physical Constants

Summary of the 1986 Recommended Values
of the Fundamental Physical Constants

Quantity	Symbol	Value	Units	Relative Uncertainty (ppm)
Speed of light in vacuum	c	299 792 458	ms^{-1}	(exact)
Permeability of vacuum	μ_o	$4\pi \times 10^{-7}$	$N\ A^{-2}$	
		= 12.566 370614 …	$10^{-7}\ N\ A^{-2}$	(exact)
Permittivity of vacuum	ε_o	$1/\mu_o c^2$		
		= 8.854 187 817 …	$10^{-12}\ F\ m^{-1}$	(exact)
Newtonian constant of gravitation	G	6.672 59(85)	$10^{-11}\ m^3\ kg^{-1}\ s^{-2}$	128
Planck constant	h	6.626 0755(40)	10^{-34} J s	0.60
$h/2\pi$	$\hbar$	1.054 572 66(63)	10^{-34} J s	0.60
Elementary charge	e	1.602 177 33(49)	10^{-19} C	0.30
Magnetic flux quantum, $h/2e$	Φ_o	2.067 834 61(61)	10^{-15} Wb	0.30
Electron mass	m_e	9.109 3897(54)	10^{-31} kg	0.59
Proton mass	m_p	1.672 6231(10)	10^{-27} kg	0.59
Proton–electron mass ratio	$m_p jm_e$	1836.152701(37)		0.020
Fine-structure constant, $\mu_o ce^2/2h$	α	7.297 353 08(33)	10^{-3}	0.045
Inverse fine-structure constant	α^{-1}	137 035 9895(61)		0.045
Rydberg constant, $m_e c\alpha^2/2h$	R_∞	10 973 731.534(13)	m^{-1}	0.0012
Avogadro constant	N_A, L	6.022 1367(36)	$10^{23}\ mol^{-1}$	0.59
Faraday constant, $N_A e$	F	96 485.309(29)	$C\ mol^{-1}$	0.30
Molar gas constant	R	8.314 510(70)	$J\ mol^{-1}\ K^{-1}$	8.4
Boltzmann constant, R/N_A	k	1.380 658(12)	$10^{-23}\ J\ K^{-1}$	8.5
Stafan–Boltzmann constant, $(\pi^2/60)k^4/\hbar^3 c^2$	σ	5.670 51(19)	$10^{-8}\ W\ m^{-2}\ K^{-4}$	34
Non-SI units used with SI				
Electronvolt, $(e/C)J = \{e\}J$	eV	1.602 17733(40)	10^{-19} J	0.30
(Unified) atomic mass unit, $1\ u = m_u = 1/12m(^{12}C)$	u	1.660 5402(10)	10^{-27} kg	0.59

Note: An abbreviated list of the fundamental constants of physics and chemistry based on a least-squares adjustment with 17 degrees of freedom. The digits in parentheses are the one-standard-deviation uncertainty in the last digits of the given value. Since the uncertainties of many entries are correlated, the full covariance matrix must be used in evaluating the uncertainties of quantities computed from them.

PERIODIC TABLE OF THE ELEMENTS

New Notation
Previous IUPAC Form
CAS Version

Key to Chart

Atomic Number →	50 +2 +4	← Oxidation States
Symbol →	Sn	
1995 Atomic Weight →	118.710 -18-18-4	← Electron Configuration

1 Group IA	2 IIA	3 IIIA IIIB	4 IVA IVB	5 VA VB	6 VIA VIB	7 VIIA VIIB	8	9 VIIIA VIII	10	11 IB IB	12 IIB IIB	13 IIIB IIIA	14 IVB IVA	15 VB VA	16 VIB VIA	17 VIIB VIIA	18 VIIIA	Shell
1 H +1 -1 1.00794 1																	2 He 0 4.002602 2	K
3 Li +1 6.941 2-1	4 Be +2 9.012182 2-2											5 B +3 10.811 2-3	6 C +2 +4 -4 12.0107 2-4	7 N +1 +2 +3 +4 +5 -1 -2 -3 14.00674 2-5	8 O -2 15.9994 2-6	9 F -1 18.9984032 2-7	10 Ne 0 20.1797 2-8	K-L
11 Na +1 22.989770 2-8-1	12 Mg +2 24.3050 2-8-2											13 Al +3 26.981538 2-8-3	14 Si +2 +4 -4 28.0855 2-8-4	15 P +3 +5 -3 30.973761 2-8-5	16 S +4 +6 -2 32.066 2-8-6	17 Cl +1 +5 +7 -1 35.4527 2-8-7	18 Ar 0 39.948 2-8-8	K-L-M
19 K +1 39.0983 -8-8-1	20 Ca +2 40.078 -8-8-2	21 Sc +3 44.955910 -8-9-2	22 Ti +2 +3 +4 47.867 -8-10-2	23 V +2 +3 +4 +5 50.9415 -8-11-2	24 Cr +2 +3 +6 51.9961 -8-13-1	25 Mn +2 +3 +4 +7 54.938049 -8-13-2	26 Fe +2 +3 55.845 -8-13-2	27 Co +2 +3 58.933200 -8-15-2	28 Ni +2 +3 58.6934 -8-16-2	29 Cu +1 +2 63.546 -8-18-1	30 Zn +2 65.39 -8-18-2	31 Ga +3 69.723 -8-18-3	32 Ge +2 +4 72.61 -8-18-4	33 As +3 +5 -3 74.92160 -8-18-5	34 Se +4 +6 -2 78.96 -8-18-6	35 Br +1 +5 -1 79.904 -8-18-7	36 Kr 0 83.80 -8-18-8	-L-M-N
37 Rb +1 85.4678 -18-8-1	38 Sr +2 87.62 -18-8-2	39 Y +3 88.90585 -18-9-2	40 Zr +4 91.224 -18-10-2	41 Nb +3 +5 92.90638 -18-12-1	42 Mo +6 95.94 -18-13-1	43 Tc +4 +6 +7 (98) -18-13-2	44 Ru +3 101.07 -18-15-1	45 Rh +3 102.90550 -18-16-1	46 Pd +2 +3 106.42 -18-18-0	47 Ag +1 107.8682 -18-18-1	48 Cd +2 112.411 -18-18-2	49 In +3 114.818 -18-18-3	50 Sn +2 +4 118.710 -18-18 -4	51 Sb +3 +5 -3 121.760 -18-18-5	52 Te +4 +6 -2 127.60 -18-18-6	53 I +1 +5 +7 -1 126.90447 -18-18-7	54 Xe 0 131.29 -18-18-8	-M-N-O
55 Cs +1 132.90545 -18-8-1	56 Ba +2 137.327 -18-8-2	57* La +3 138.9055 -18-9-2	72 Hf +4 178.49 -32-10-2	73 Ta +5 180.9479 -32-11-2	74 W +6 183.84 -32-12-2	75 Re +4 +6 +7 186.207 -32-13-2	76 Os +3 +4 190.23 -32-14-2	77 Ir +3 +4 192.217 -32-15-2	78 Pt +2 +4 195.078 -32-17-1	79 Au +1 +3 196.96655 -32-18-1	80 Hg +1 +2 200.59 -32-18-2	81 Tl +1 +3 204.3833 -32-18-3	82 Pb +2 +4 207.2 -32-18-4	83 Bi +3 +5 208.98038 -32-18-5	84 Po +2 +4 (209) -32-18-6	85 At (210) -32-18-7	86 Rn 0 (222) -32-18-8	-N-O-P
87 Fr +1 (223) -18-8-1	88 Ra +2 (226) -18-8-2	89** Ac +3 (227) -18-9-2	104 Rf +4 (261) -32-10-2	105 Db (262) -32-11-2	106 Sg (266) -32-12-2	107 Bh (264) -32-13-2	108 Hs (269) -32-14-2	109 Mt (268) -32-15-2	110 Uun (271) -32-16-2	111 Uuu (272)	112 Uub							-O-P-Q

* Lanthanides	58 Ce +3 +4 140.116 -19-9-2	59 Pr +3 140.90765 -21-8-2	60 Nd +3 144.24 -22-8-2	61 Pm +3 (145) -23-8-2	62 Sm +2 +3 150.36 -24-8-2	63 Eu +2 +3 151.964 -25-8-2	64 Gd +3 157 .25 -25-9-2	65 Tb +3 158.92534 -27-8-2	66 Dy +3 162.50 -28-8-2	67 Ho +3 164.93032 -29-8-2	68 Er +3 167.26 -30-8-2	69 Tm +3 168.93421 -31-8-2	70 Yb +2 +3 173.04 -32-8-2	71 Lu +3 174.967 -32-9-2	-N-O-P
** Actinides	90 Th +4 232.0381 -18-10-2	91 Pa +5 +4 231.03588 -20-9-2	92 U +3 +4 +5 +6 238.0289 -21-9-2	93 Np +3 +4 +5 +6 (237) -22-9-2	94 Pu +3 +4 +5 +6 (244) -24-8-2	95 Am +3 +4 +5 +6 (243) -25-8-2	96 Cm +3 (247) -25-9-2	97 Bk +3 +4 (247) -27-8-2	98 Cf +3 (251) -28-8-2	99 Es +3 (252) -29-8-2	100 Fm +3 (257) -30-8-2	101 Md +2 +3 (258) -31-8-2	102 No +2 +3 (259) -32-8-2	103 Lr +3 (262) -32-9-2	-O-P-Q

The new IUPAC format numbers the groups from 1 to 18. The previous IUPAC numbering system and the system used by Chemical Abstracts Service (CAS) are also shown. For radioactive elements that do not occur in nature, the mass number of the most stable isotope is given in parentheses.

References
1. G. J. Leigh, Editor, *Nomenclature of Inorganic Chemistry*, Blackwell Scientific Publications, Oxford, 1990.
2. *Chemical and Engineering News*, 63(5), 27, 1985.
3. Atomic Weights of the Elements, 1995, *Pure & Appl. Chem.*, 68, 2339, 1996.

Electrical Resistivity

Electrical Resistivity of Pure Metals

The first part of this table gives the electrical resistivity, in units of $10^{-8}\ \Omega$ m, for 28 common metallic elements as a function of temperature. The data refer to polycrystalline samples. The number of significant figures indicates the accuracy of the values. However, at low temperatures (especially below 50 K) the electrical resistivity is extremely sensitive to sample purity. Thus the low-temperature values refer to samples of specified purity and treatment.

The second part of the table gives resistivity values in the neighborhood of room temperature for other metallic elements that have not been studied over an extended temperature range.

Electrical Resistivity in $10^{-8}\ \Omega$ m

T/K	Aluminum	Barium	Beryllium	Calcium	Cesium	Chromium	Copper
1	0.000100	0.081	0.0332	0.045	0.0026		0.00200
10	0.000193	0.189	0.0332	0.047	0.243		0.00202
20	0.000755	0.94	0.0336	0.060	0.86	0.00280	
40	0.0181	2.91	0.0367	0.175	1.99		0.0239
60	0.0959	4.86	0.067	0.40	3.07		0.0971
80	0.245	6.83	0.075	0.65	4.16		0.215
100	0.442	8.85	0.133	0.91	5.28	1.6	0.348
150	1.006	14.3	0.510	1.56	8.43	4.5	0.699
200	1.587	20.2	1.29	2.19	12.2	7.7	1.046
273	2.417	30.2	3.02	3.11	18.7	11.8	1.543
293	2.650	33.2	3.56	3.36	20.5	12.5	1.678
298	2.709	34.0	3.70	3.42	20.8	12.6	1.712
300	2.733	34.3	3.76	3.45	21.0	12.7	1.725
400	3.87	51.4	6.76	4.7		15.8	2.402
500	4.99	72.4	9.9	6.0		20.1	3.090
600	6.13	98.2	13.2	7.3		24.7	3.792
700	7.35	130	16.5	8.7		29.5	4.514
800	8.70	168	20.0	10.0	34.6	5.262	
900	10.18	216	23.7	11.4		39.9	6.041

T/K	Gold	Hafnium	Iron	Lead	Lithium	Magnesium	Manganese
1	0.0220	1.00	0.0225		0.007	0.0062	7.02
10	0.0226	1.00	0.0238		0.008	0.0069	18.9
20	0.035	1.11	0.0287		0.012	0.0123	54
40	0.141	2.52	0.0758		0.074	0.074	116
60	0.308	4.53	0.271		0.345	0.261	131
80	0.481	6.75	0.693	4.9	1.00	0.557	132
100	0.650	9.12	1.28	6.4	1.73	0.91	132
150	1.061	15.0	3.15	9.9	3.72	1.84	136
200	1.462	21.0	5.20	13.6	5.71	2.75	139
273	2.051	30.4	8.57	19.2	8.53	4.05	143
293	2.214	33.1	9.61	20.8	9.28	4.39	144
298	2.255	33.7	9.87	21.1	9.47	4.48	144
300	2.271	34.0	9.98	21.3	9.55	4.51	144
400	3.107	48.1	16.1	29.6	13.4	6.19	147
500	3.97	63.1	23.7	38.3		7.86	149
600	4.87	78.5	32.9			9.52	151
700	5.82		44.0			11.2	152
800	6.81		57.1			12.8	
900	7.86					14.4	

(continued)

T/K	Molybdenum	Nickel	Palladium	Platinum	Potassium	Rubidium	Silver
1	0.00070	0.0032	0.0200	0.002	0.0008	0.0131	0.00100
10	0.00089	0.0057	0.0242	0.0154	0.0160	0.109	0.00115
20	0.00261	0.0140	0.0563	0.0484	0.117	0.444	0.0042
40	0.0457	0.068	0.334	0.409	0.480	1.21	0.0539
60	0.206	0.242	0.938	1.107	0.90	1.94	0.162
80	0.482	0.545	1.75	1.922	1.34	2.65	0.289
100	0.858	0.96	2.62	2.755	1.79	3.36	0.418
150	1.99	2.21	4.80	4.76	2.99	5.27	0.726
200	3.13	3.67	6.88	6.77	4.26	7.49	1.029
273	4.85	6.16	9.78	9.6	6.49	11.5	1.467
293	5.34	6.93	10.54	10.5	7.20	12.8	1.587
298	5.47	7.12	10.73	10.7	7.39	13.1	1.617
300	5.52	7.20	10.80	10.8	7.47	13.3	1.629
400	8.02	11.8	14.48	14.6			2.241
500	10.6	17.7	17.94	18.3			2.87
600	13.1	25.5	21.2	21.9			3.53
700	15.8	32.1	24.2	25.4			4.21
800	18.4	35.5	27.1	28.7			4.91
900	21.2	38.6	29.4	32.0			5.64

T/K	Sodium	Strontium	Tantalum	Tungsten	Vanadium	Zinc	Zirconium
1	0.0009	0.80	0.10	0.00016		0.0100	0.250
10	0.0015	0.80	0.102	0.000137	0.0145	0.0112	0.253
20	0.016	0.92	0.146	0.00196	0.039	0.0387	0.357
40	0.172	1.70	0.751	0.0544	0.304	0.306	1.44
60	0.447	2.68	1.65	0.266	1.11	0.715	3.75
80	0.80	3.64	2.62	0.606	2.41	1.15	6.64
100	1.16	4.58	3.64	1.02	4.01	1.60	9.79
150	2.03	6.84	6.19	2.09	8.2	2.71	17.8
200	2.89	9.04	8.66	3.18	12.4	3.83	26.3
273	4.33	12.3	12.2	4.82	18.1	5.46	38.8
293	4.77	13.2	13.1	5.28	19.7	5.90	42.1
298	4.88	13.4	13.4	5.39	20.1	6.01	42.9
300	4.93	13.5	13.5	5.44	20.2	6.06	43.3
400		17.8	18.2	7.83	28.0	8.37	60.3
500		22.2	22.9	10.3	34.8	10.82	76.5
600		26.7	27.4	13.0	41.1	13.49	91.5
700		31.2	31.8	15.7	47.2		104.2
800		35.6	35.9	18.6	53.1		114.9
900			40.1	21.5	58.7		123.1

Electrical Resistivity of Pure Metals (continued)

Element	T/K	Electrical Resistivity 10^{-8} Ω m
Antimony	273	39
Bismuth	273	107
Cadmium	273	6.8
Cerium	290–300	82.8
Cobalt	273	5.6
Dysprosium	290–300	92.6
Erbium	290–300	86.0
Europium	290–300	90.0
Gadolinium	290–300	131
Gallium	273	13.6
Holmium	290–300	81.4
Indium	273	8.0
Iridium	273	4.7
Lanthanum	290–300	61.5
Lutetium	290–300	58.2
Mercury	273	94.1
Neodymium	290–300	64.3
Niobium	273	15.2
Osmium	273	8.1
Polonium	273	40
Praseodymium	290–300	70.0
Promethium	290–300	75
Protactinium	273	17.7
Rhenium	273	17.2
Rhodium	273	4.3
Ruthenium	273	7.1
Samrium	290–300	94.0
Scandium	290–300	56.2
Terbium	290–300	115
Thallium	273	15
Thorium	273	14.7
Thulium	290–300	67.6
Tin	273	11.5
Titanium	273	39
Uranium	273	28
Ytterbium	290–300	25.0
Yttrium	290–300	59.6

Electrical Resistivity of Selected Alloys

Values of the resistivity are given in units of 10^{-8} Ω m. General comments in the preceding table for pure metals also apply here.

Alloy—Aluminum-Copper

Wt % Al	273 K	293 K	300 K	350 K	400 K
99[a]	2.51	2.74	2.82	3.38	3.95
95[a]	2.88	3.10	3.18	3.75	4.33
90[b]	3.36	3.59	3.67	4.25	4.86
85[b]	3.87	4.10	4.19	4.79	5.42
80[b]	4.33	4.58	4.67	5.31	5.99
70[b]	5.03	5.31	5.41	6.16	6.94
60[b]	5.56	5.88	5.99	6.77	7.63
50[b]	6.22	6.55	6.67	7.55	8.52
40[c]	7.57	7.96	8.10	9.12	10.2
30[c]	11.2	11.8	12.0	13.5	15.2
25[f]	16.3[aa]	17.2	17.6	19.8	22.2
15[h]	—	12.3	—	—	—
19[g]	10.8[aa]	11.0	11.1	11.7	12.3
5[e]	9.43	9.61	9.68	10.2	10.7
1[b]	4.46	4.60	4.65	5.00	5.37

Alloy—Aluminum-Magnesium

Wt % Al	273 K	293 K	300 K	350 K	400 K
99[c]	2.96	3.18	3.26	3.82	4.39
95[c]	5.05	5.28	5.36	5.93	6.51
90[c]	7.52	7.76	7.85	8.43	9.02
85	—	—	—	—	—
80	—	—	—	—	—
70	—	—	—	—	—
60	—	—	—	—	—
50	—	—	—	—	—
40	—	—	—	—	—
30	—	—	—	—	—
25	—	—	—	—	—
15	—	—	—	—	—
10[b]	17.1	17.4	17.6	18.4	19.2
5[b]	13.1	13.4	13.5	14.3	15.2
1[a]	5.92	6.25	6.37	7.20	8.03

Alloy—Copper-Gold

Wt % Cu	273 K	293 K	300 K	350 K	400 K
99[c]	1.73	1.86[aa]	1.91[aa]	2.24[aa]	2.58[aa]
95[c]	2.41	2.54[aa]	2.59[aa]	2.92[aa]	3.26[aa]
90[c]	3.29	4.42[aa]	3.46[aa]	3.79[aa]	4.12[aa]
85[c]	4.20	4.33	4.38[aa]	4.71[aa]	5.05[aa]
80[c]	5.15	5.28	5.32	5.65	5.99
70[c]	7.12	7.25	7.30	7.64	7.99
60[c]	9.18	9.13	9.36	9.70	10.05
50[c]	11.07	11.20	11.25	11.60	11.94
40[c]	12.70	12.85	12.90[aa]	13.27[aa]	13.65[aa]
30[c]	13.77	13.93	13.99[aa]	14.38[aa]	14.78[aa]
25[c]	13.93	14.09	14.14	14.54	14.94
15[c]	12.75	12.91	12.96[aa]	13.36[aa]	13.77
10[c]	10.70	10.86	10.91	11.31	11.72
5[c]	7.25	7.41[aa]	7.46	7.87	8.28
1[c]	3.40	3.57	3.62	4.03	4.45

Alloy—Copper-Nickel

Wt % Cu	273 K	293 K	300 K	350 K	400 K
99[c]	2.71	2.85	2.91	3.27	3.62
95[c]	7.60	7.71	7.82	8.22	8.62
90[c]	13.69	13.89	13.96	14.40	14.81
85[c]	19.63	19.83	19.90	2032	20.70
80[c]	25.46	25.66	25.72	26.12[aa]	26.44[aa]
70[i]	36.67	36.72	36.76	36.85	36.89
60[i]	45.43	45.38	45.35	45.20	45.01
50[i]	50.19	50.05	50.01	49.73	49.50
40[c]	47.42	47.73	47.82	48.28	48.49
30[i]	40.19	41.79	42.34	44.51	45.40
25[c]	33.46	35.11	35.69	39.67[aa]	42.81[aa]
15[c]	22.00	23.35	23.85	27.60	31.38
10[c]	16.65	17.82	18.26	21.51	25.19
5[c]	11.49	12.50	12.90	15.69	18.78
1[c]	7.23	8.08	8.37	10.63[aa]	13.18[aa]

Alloy—Copper-Palladium

Wt % Cu	273 K	293 K	300 K	350 K	400 K
99[c]	2.10	2.23	2.27	2.59	2.92
95[c]	4.21	4.35	4.40	4.74	5.08
90[c]	6.89	7.03	7.08	7.41	7.74
85[c]	9.48	9.61	9.66	10.01	10.36
80[c]	11.99	12.12	12.16	12.51[aa]	12.87
70[c]	16.87	17.01	17.06	17.41	17.78
60[c]	21.73	21.87	21.92	22.30	22.69
50[c]	27.62	27.79	27.86	28.25	28.64
40[c]	35.31	35.51	35.57	36.03	36.47
30[c]	46.50	46.66	46.71	47.11	47.47
25[c]	46.25	46.45	46.52	46.99[aa]	47.43[aa]
15[c]	36.52	36.99	37.16	38.28	39.35
10[c]	28.90	29.51	29.73	31.19[aa]	32.56[aa]
5[c]	20.00	20.75	21.02	22.84[aa]	24.54[aa]
1[c]	11.90	12.67	12.93[aa]	14.82[aa]	16.68[aa]

Alloy—Copper-Zinc

Wt % Cu	273 K	293 K	300 K	350 K	400 K
99[b]	1.84	1.97	2.02	2.36	2.71
95[b]	2.78	2.92	2.97	3.33	3.69
90[b]	3.66	3.81	3.86	4.25	4.63
85[b]	4.37	4.54	4.60	5.02	5.44
80[b]	5.01	5.19	5.26	5.71	6.17
70[b]	5.87	6.08	6.15	6.67	7.19
60	—	—	—	—	—
50	—	—	—	—	—
40	—	—	—	—	—
30	—	—	—	—	—
25	—	—	—	—	—
15	—	—	—	—	—
10	—	—	—	—	—
5	—	—	—	—	—
1	—	—	—	—	—

Alloy—Gold-Palladium

Wt % Au	273 K	293 K	300 K	350 K	400 K
99[c]	2.69	2.86	2.91	3.32	3.73
95[c]	5.21	5.35	5.41	5.79	6.17
90[i]	8.01	8.17	8.22	8.56	8.93
85[b]	10.50[aa]	10.66	10.72[aa]	11.100[aa]	11.48[aa]
80[b]	12.75	12.93	12.99	13.45	13.93
70[c]	18.23	18.46	18.54	19.10	19.67
60[b]	26.70	26.94	27.01	27.63[aa]	28.23[aa]
50[a]	27.23	27.63	27.76	28.64[aa]	29.42[aa]
40[a]	24.65	25.23	25.42	26.74	27.95
30[b]	20.82	21.49	21.72	23.35	24.92
25[b]	18.86	19.53	19.77	21.51	23.19
15[a]	15.08	15.77	16.01	17.80	19.61
10[a]	13.25	13.95	14.20[aa]	16.00[aa]	17.81[aa]
5[a]	11.49[aa]	12.21	12.46[aa]	14.26[aa]	16.07[aa]
1[a]	10.07	10.85[aa]	11.12[aa]	12.99[aa]	14.80[aa]

Alloy—Iron-Nickel

Wt % Fe	273 K	293 K	300 K	350 K	400 K
99[a]	10.9	12.0	12.4	—	18.7
95[c]	18.7	19.9	20.2	—	26.8
90[c]	24.2	25.5	25.9	—	33.2
85[c]	27.8	29.2	29.7	—	37.3
80[c]	30.1	31.6	32.2	—	40.0
70[b]	32.3	33.9	34.4	—	42.4
60[c]	53.8	57.1	58.2	—	73.9
50[d]	28.4	30.6	31.4	—	43.7
40[d]	19.6	21.6	22.5	—	34.0
30[c]	15.3	17.1	17.7	—	27.4
25[b]	14.3	15.9	16.4	—	25.1
15[c]	12.6	13.8	14.2	—	21.1
10[c]	11.4	12.5	12.9	—	18.9
5[c]	9.66	10.6	10.9	—	16.1[aa]
1[b]	7.17	7.94	8.12	—	12.8

Alloy—Gold-Silver

Wt % Au	273 K	293 K	300 K	350 K	400 K
99[b]	2.58	2.75	2.80[aa]	3.22[aa]	3.63[aa]
95[a]	4.58	4.74	4.79	5.19	5.59
90[j]	6.57	6.73	6.78	7.19	7.58
85[j]	8.14	8.30	8.36[aa]	8.75	9.15
80[j]	9.34	9.50	9.55	9.94	10.33
70[j]	10.70	10.86	10.91	11.29	11.68[aa]
60[j]	10.92	11.07	11.12	11.50	11.87
50[j]	10.23	10.37	10.42	10.78	11.14
40[j]	8.92	9.06	9.11	9.46[aa]	9.81
30[a]	7.34	7.47	7.52	7.85	8.19
25[a]	6.46	6.59	6.63	6.96	7.30[aa]
15[a]	4.55	4.67	4.72	5.03	5.34
10[a]	3.54	3.66	3.71	4.00	4.31
5[i]	2.52	2.64[aa]	2.68[aa]	2.96[aa]	3.25[aa]
1[b]	1.69	1.80	1.84[aa]	2.12[aa]	2.42[aa]

Alloy—Silver-Palladium

Wt % Ag	273 K	293 K	300 K	350 K	400 K
99[b]	1.891	2.007	2.049	2.35	2.66
95[b]	3.58	3.70	3.74	4.04	4.34
90[b]	5.82	5.94	5.98	6.28	6.59
85[k]	7.92[aa]	8.04[aa]	8.08	8.38[aa]	8.68[aa]
80[k]	10.01	10.13	10.17	10.47	10.78
70[k]	14.53	14.65	14.69	14.99	15.30
60[i]	20.9	21.1	21.2	21.6	22.0
50[k]	31.2	31.4	31.5	32.0	32.4
40[m]	42.2	42.2	42.2	42.3	42.3
30[b]	40.4	40.6	40.7	41.3	41.7
25[k]	36.67[aa]	37.06	37.19	38.1[aa]	38.8[aa]
15[i]	27.08[aa]	26.68[aa]	27.89[aa]	29.3[aa]	30.6[aa]
10[i]	21.69	22.39	22.63	24.3	25.9
5[b]	15.98	16.72	16.98	18.8[aa]	20.5[aa]
1[a]	11.06	11.82	12.08[aa]	13.92[aa]	15.70[aa]

[a] Uncertainty in resistivity is ± 2%.
[b] Uncertainty in resistivity is ± 3%.
[c] Uncertainty in resistivity is ± 5%.
[d] Uncertainty in resistivity is ± 7% below 300 K and ± 5% at 300 and 400 K.
[e] Uncertainty in resistivity is ± 7%.
[f] Uncertainty in resistivity is ± 8%.
[g] Uncertainty in resistivity is ± 10%.
[h] Uncertainty in resistivity is ± 12%.
[i] Uncertainty in resistivity is ± 4%.
[j] Uncertainty in resistivity is ± 1%.
[k] Uncertainty in resistivity is ± 3% up to 300 K and ± 4% above 300 K.
[m] Uncertainty in resistivity is ± 2% up to 300 K and ± 4% above 300 K.
[a] Crystal usually a mixture of α-hep and fcc lattice.
[aa] In temperature range where no experimental data are available.

Resistivity of Selected Ceramics (Listed by Ceramic)

Ceramic	Resistivity (Ω-cm)
Borides	
Chromium diboride (CrB_2)	21×10^{-6}
Hafnium diboride (HfB_2)	$10–12 \times 10^{-6}$ at room temp.
Tantalum diboride (TaB_2)	68×10^{-6}
Titanium diboride (TiB_2) (polycrystalline)	
85% dense	$26.5–28.4 \times 10^{-6}$ at room temp.
85% dense	9.0×10^{-6} at room temp.
100% dense, extrapolated values	$8.7–14.1 \times 10^{-6}$ at room temp.
	3.7×10^{-6} at liquid air temp.
Titanium diboride (TiB_2) (monocrystalline)	
Crystal length 5 cm, 39 deg. and 59 deg. orientation with respect to growth axis	$6.6 \pm 0.2 \times 10^{-6}$ at room temp.
Crystal length 1.5 cm, 16.5 deg. and 90 deg. orientation with respect to growth axis	$6.7 \pm 0.2 \times 10^{-6}$ at room temp.
Zirconium diboride (ZrB_2)	9.2×10^{-6} at 20°C
	1.8×10^{-6} at liquid air temp.
Carbides: boron carbide (B_4C)	0.3–0.8

Dielectric Constants

Dielectric Constants of Solids

These data refer to temperatures in the range 17–22°C.

Material	Freq. (Hz)	Dielectric Constant	Material	Freq. (Hz)	Dielectric Constant
Acetamide	4×10^8	4.0	Diphenylmethane	4×10^8	2.7
Acetanilide	—	2.9	Dolomite ⊥ optic axis	10^8	8.0
Acetic acid (2°C)	4×10^8	4.1	Dolomite ∥	10^8	6.8
Aluminum oleate	4×10^8	2.40	Ferrous oxide (15°C)	10^8	14.2
Ammonium bromide	10^8	7.1	Iodine	10^8	4
Ammonium chloride	10^8	7.0	Lead acetate	10^8	2.6
Antimony trichloride	10^8	5.34	Lead carbonate (15°C)	10^8	18.6
Apatite ⊥ optic axis	3×10^8	9.50	Lead chloride	10^8	4.2
Apatite ∥ optic axis	3×10^8	7.41	Lead monoxide (15°C)	10^8	25.9
Asphalt	$<3 \times 10^4$	2.68	Lead nitrate	6×10^7	37.7
Barium chloride (anhyd.)	6×10^7	11.4	Lead oleate	4×10^8	3.27
Barium chloride ($2H_3O$)	6×10^7	9.4	Lead sulfate	10^4	14.3
Barium nitrate	6×10^7	5.9	Lead sulfide (15°C)	16^8	17.9
Barium sulfate (15°C)	10^8	11.40	Malachite (mean)	10^{12}	7.2
Beryl ⊥ optic axis	10^4	7.02	Mercuric chloride	10^8	3.2
Beryl ∥ optic axis	10^4	6.08	Mercurous chloride	10^8	9.4
Calcite ⊥ optic axis	10^4	8.5	Naphthalene	4×10^8	2.52
Calcite ∥ optic axis	10^4	8.0	Phenanthrene	4×10^8	2.80
Calcium carbonate	10^4	6.14	Phenol (10°C)	4×10^8	4.3
Calcium fluoride	10^4	7.36	Phosphorus, red	10^8	4.1
Calcium sulfate ($2H_2O$)	10^4	5.66	Phosphorus, yellow	10^8	3.6
Cassiterite ⊥ optic axis	10^{12}	23.4	Potassium aluminum sulfate	10^8	3.8
Cassiterite ∥ optic axis	10^{12}	24	Potassium carbonate (15°C)	10^8	5.6
d-Cocaine	5×10^8	3.10	Potassium chlorate	6×10^7	5.1
Cupric oleate	4×10^8	2.80	Potassium chloride	10^4	5.03
Cupric oxide (15°C)	10^8	18.1	Potassium chromate	6×10^7	7.3
Cupric sulfate (anhyd.)	6×10^7	10.3	Potassium iodide	6×10^7	5.6
Cupric sulfate ($5H_2O$)	6×10^7	7.8	Potassium nitrate	6×10^7	5.0
Diamond	10^8	5.5	Potassium sulfate	6×10^7	5.9

Material	Freq. (Hz)	Dielectric Constant	Material	Freq. (Hz)	Dielectric Constant
Quartz ⊥ optic axis	3×10^7	4.34	Sodium carbonate ($10H_2O$)	6×10^7	5.3
Quartz ∥ optic axis	3×10^7	4.27	Sodium chloride	10^4	6.12
Resorcinol	4×10^8	3.2	Sodium nitrate	—	5.2
Ruby ⊥ optic axis	10^4	13.27	Sodium oleate	4×10^8	2.75
Ruby ∥ optic axis	10^4	11.28	Sodium perchlorate	6×10^7	5.4
Rutile ⊥ optic axis	10^8	86	Sucrose (mean)	3×10^8	3.32
Rutile ∥ optic axis	10^8	170	Sulfur (mean)	—	4.0
Selenium	10^8	6.6	Thallium chloride	10^4	46.9
Silver bromide	10^4	12.2	*p*-Toluidine	4×10^8	3.0
Silver chloride	10^4	11.2	Tourmaline ⊥ optic axis	10^4	7.10
Silver cyanide	10^4	5.6	Tourmaline ∥ optic axis	10^4	6.3
Smithsonite ⊥ optic axis	10^{12}	9.3	Urea	4×10^8	3.5
Smithsonite ∥ optic axis	10^{10}	9.4	Zircon ⊥, ∥	10^8	12
Sodium carbonate (anhyd.)	6×10^7	8.4			

Dielectric Constants of Ceramics

Material	Dielectric Constant 10^4 Hz	Dielectric strength Volts/mil	Volume Resistivity Ohm-cm (23°C)	Loss Factor[a]
Alumina	4.5–8.4	40–160	10^{11}–10^{14}	0.0002–0.01
Corderite	4.5–5.4	40–250	10^{12}–10^{14}	0.004–0.012
Forsterite	6.2	240	10^{14}	0.0004
Porcelain (dry process)	6.0–8.0	40–240	10^{12}–10^{14}	0.0003–0.02
Porcelain (wet process)	6.0–7.0	90–400	10^{12}–10^{14}	0.006–0.01
Porcelain, zircon	7.1–10.5	250–400	10^{13}–10^{15}	0.0002–0.008
Steatite	5.5–7.5	200–400	10^{13}–10^{15}	0.0002–0.004
Titanates (Ba, Sr, Ca, Mg, and Pb)	15–12.000	50–300	10^{8}–10^{13}	0.0001–0.02
Titanium dioxide	14–110	100–210	10^{13}–10^{18}	0.0002–0.005

Dielectric Constants of Glasses

Type	Dielectric Constant At 100 MHz (20°C)	Volume Resistivity (350°C megohm-cm)	Loss Factor[a]
Corning 0010	6.32	10	0.015
Corning 0080	6.75	0.13	0.058
Corning 0120	6.65	100	0.012
Pyrex 1710	6.00	2,500	0.025
Pyrex 3320	4.71	—	0.019
Pyrex 7040	4.65	80	0.013
Pyrex 7050	4.77	16	0.017
Pyrex 7052	5.07	25	0.019
Pyrex 7060	4.70	13	0.018
Pyrex 7070	4.00	1,300	0.0048
Vycor 7230	3.83	—	0.0061
Pyrex 7720	4.50	16	0.014
Pyrex 7740	5.00	4	0.040
Pyrex 7750	4.28	50	0.011
Pyrex 7760	4.50	50	0.0081
Vycor 7900	3.9	130	0.0023
Vycor 7910	3.8	1,600	0.00091
Vycor 7911	3.8	4,000	0.00072
Corning 8870	9.5	5,000	0.0085
G. E. Clear (silica glass)	3.81	4,000–30,000	0.00038
Quartz (fused)	3.75 4.1 (1 MHz)	—	0.0002 (1 MHz)

[a] Power factor × dielectric constant equals loss factor.

Properties of Semiconductors

H. Mike Harris *Georgia Tech Research Institute*

Semiconducting Properties of Selected Materials

Substance	Minimum Energy Gap (eV) R.T.	Minimum Energy Gap (eV) 0 K	$\frac{dE_g}{dT}$ $\times 10^4$ eV/°C	$\frac{dE_g}{dP}$ $\times 10^6$ eV·cm²/kg	Density of States eleCtron Effective Mass m_{d_n} (m_o)	Electron Mobility and Temperature Dependence μ_n (cm²/V·s)	Electron Mobility and Temperature Dependence $-x$	Density of States Hole Effec-Tive Mass m_{d_p} (m_o)	Hole Mobility and Temperature Dependence μ_p (cm²/V·s)	Hole Mobility and Temperature Dependence $-x$
Si	1.107	1.153	–2.3	–2.0	1.1	1,900	2.6	0.56	500	2.3
Ge	0.67	0.744	–3.7	±7.3	0.55	3,80	1.66	0.3	1,820	2.33
αSn	0.08	0.094	–0.5		0.02	2,500	1.65	0.3	2,400	2.0
Te	0.33				0.68	1,100		0.19	560	
III–V Compounds										
AlAs	2.2	2.3				1,200			420	
AlSb	1.6	1.7	–3.5	–1.6	0.09	2..	1.5	0.4	500	1.8
GaP	2.24	2.40	–5.4	–1.7	0.35	300	1.5	0.5	150	1.5
GaAs	1.35	1.53	–5.0	+9.4	0.068	9,000	1.0	0.5	500	2.1
GaSb	0.67	0.78	–3.5	+12	0.050	5,000	2.0	0.23	1,400	0.9
InP	1.27	1.41	–4.6	+4.6	0.067	5,000	2.0		200	2.4
InAs	0.36	0.43	–2.8	+8	0.022	33,000	1.2	0.41	460	2.3
InSb	0.165	0.23	–2.8	+15	0.014	78,000	1.6	0.4	750	2.1
II–VI Compounds										
ZnO	3.2		–9.5	+0.6	0.38	180	1.5			
ZnS	3.54		–5.3	+5.7		180			5 (400°C)	
ZnSe	2.58	2.80	–7.2	+6		540			28	
ZnTe	2.26			+6		340			100	
CdO	2.5 ± 0.1		–6		0.1	120				
CdS	2.42		–5	+3.3	0.165	400		0.8		
CdSe	1.74	1.85	–4.6		0.13	650	1.0	0.6		
CdTe	1.44	1.56	–4.1	+8	0.14	1,200		0.35	50	
HgSe	0.30				0.030	20,000	2.0			
HgTe	0.15		–1		0.017	25,000		0.5	350	
Halite Structure Compounds										
PbS	0.37	0.28	+4		0.16	800		0.1	1,000	2.2
PbSe	0.26	0.16	+4		0.3	1,500		0.34	1,500	2.2
PbTe	0.25	0.19	+4	–7	0.21	1,600		0.14	750	2.2
Others										
ZnSb	0.50	0.56			0.15	10				1.5
CdSb	0.45	0.57	–5.4		0.15	300			2,000	1.5
Bi_2S_3	1.3					200			1,100	
Bi_2Se_3	0.27					600			675	
Bi_2Te_3	0.13		–0.95		0.58	1,200	1.68	1.07	510	1.95
Mg_2Si		0.77	–6.4		0.46	400	2.5		70	
Mg_2Ge		0.74	–9			280	2		110	
Mg_2Sn	0.21	0.33	–3.5		0.37	320			260	
Mg_3Sb_2		0.32				20			82	
Zn_3As_2	0.93					10	1.1		10	
Cd_3As_2	0.55				0.046	100,000	0.88			
GaSe	2.05		3.8						20	
GaTe	1.66	1.80	–3.6			14	–5			
InSe	1.8					9000				
TlSe	0.57		–3.9		0.3	30		0.6	20	1.5
$CdSnAs_2$	0.23				0.05	25,000	1.7			
Ga_2Te_3	1.1	1.55	–4.8							
α-In_2Te_3	1.1	1.2			0.7				50	1.1
β-In_2Te_3	1.0								5	
$Hg_5In_2Te_8$	0.5								11,000	
SnO_2									78	

Band Properties of Semiconductors

Part A. Data on Valence Bands of Semiconductors (Room Temperature)

		Band curvature effective mass (expressed as fraction of free electron mass)			
Substance	Heavy Holes	Light Holes	"Split-off" Band Holes	Energy Separation of "Split-off" Band (eV)	Measured Light Hole Mobility ($cm^2/V \cdot s$)
Semiconductors with Valence Bands Maximum at the Center of the Brillouin Zone ("F")					
Si	0.52	0.16	0.25	0.044	500
Ge	0.34	0.043	0.08	0.3	1,820
Sn	0.3				2,400
AlAs					
AlSb	0.4			0.7	550
GaP				0.13	100
GaAs	0.8	0.12	0.20	0.34	400
GaSb	0.23	0.06		0.7	1,400
InP				0.21	150
InAs	0.41	0.025	0.083	0.43	460
InSb	0.4	0.015		0.85	750
CdTe	0.35				50
HgTe	0.5				350

Semiconductors with Multiple Valence Band Maxima

		Band Curvature Effective Masses			
Substance	Number of Equivalent Valleys and Directions	Longitudinal m_L	Transverse m_T	Anisotropy $K = m_L/m_T$	Measured (light) Hole Mobility $cm^2/V \cdot s$
PbSe	4 "L" [111]	0.095	0.047	2.0	1,500
PbTe	4 "L" [111]	0.27	0.02	10	750
Bi_2Te_3	6	0.207	~0.045	4.5	515

Part B. Data on Conduction Bands of Semiconductors (Room Temperature Data)

Single Valley Semiconductors

Substance	Energy Gap (eV)	Effective Mass (m_o)	Mobility ($cm^2/V \cdot s$)
GaAs	1.35	0.067	8,500
InP	1.27	0.067	5,000
InAs	0.36	0.022	33,000
InSb	0.165	0.014	78,000
CdTe	1.44	0.11	1,000

Multivalley Semiconductors

			Band curvature effective mass		
Substance	Energy Gap	Number of equivalent valleys and direction	Longitudinal m_L	Transverse m_T	Anisotropy $K = m_L/m_T$
Si	1.107	6 in [100] "Δ"	0.90	0.192	4.7
Ge	0.67	4 in [111] at "L"	1.588	~0.0815	19.5
GaSb	0.67	as Ge	~1.0	~0.2	~5
PbSe	0.26	4 in [111] at "L"	0.085	0.05	1.7
PbTe	0.25	4 in [111] at "L"	0.21	0.029	5.5
Bi_2Te_3	0.13	6			~0.05

References

1. L. Patrick and W. J. Choyke, *Physics Review Journal*, vol. 2, p. 2255, 1977.
2. W. J. Choyke, NATO ASI Ser. E, Applied Sciences, vol. 185, and references therein, 1990.
3. A. Suzuki, A. Ogura K. Furukawa, Y. Fuji, M. Shigeta and S. Nakajima, *Journal of Applied Physics*, vol. 64, p. 2818, 1988.
4. B. W. Wessels and J. C. Gatos, *Journal of Physics Chemical Solids*, vol. 15, p. 83, 1977.
5. M. M. Anikin, A. S. Zubrilov, A. A. Lebedev, A. M. Sterlchuck, and A. E. Cherenkov, *Fitz. Tekh. Polurpovdn.*, vol. 24, p. 467, 1992.
6. *CRC Materials Science and Engineering Handbook*, p. 304.
7. D. R. Lide, Properties of Semiconductors, *CRC Handbook of Chemistry and Physics*, pp. 12–98, 1999–2000 ed. CRC Press 2000.
8. http://vshields.jpl.nasa.gov/windchime.html
9. P. T. B. Schaffer and R. G. Naum, *Journal of Optical Society of America*, vol. 59, p. 1498, 1969.
10. P. T. B. Schaffer, *Applied Optics*, vol. 10, p. 1034, 1977.
11. M. N. Yoder, *Wide Bandgap Semiconductor Materials and Devices*, vol. 43, no.10, p.1634, 1966.
12. E. L. Kern, H. W. Hamill, H. W. Deem and H. D. Sheets, *Mater. Res. Bull.*, vol. 4, p. 107, 1969.
13. D. Morelli, J. Hermans, C. Bettz, W. S. Woo, G. L. Harris and C. Taylor, *Inst. Physics Conf. Ser.*, vol. 137, p. 313–6, 1990.
14. I. I. Parafenova, Y. M. Tairov and V. F. Tsvetkov, Sov. *Physics-Semiconductors*, vol. 24, no. 2, pp. 158–61, 1990.
15. http://www.ioffe.rssi.ru/SVA/NSM/Nano/index.html
16. C.W. Garland and K. C. Parks, *Journal Applied Physics*, vol. 33, p. 759, 1962.
17. http://mems.isi.edu/mems/materials/measurements.cgi?MATTAG=galliumarsenidegaas-bulk&PAGE_SIZE=20
18. M. C. Benjamin, C. Wang, R. F. Davis and R. J. Nemanich, *Applied Phys. Letter*, vol. 64, p. 3288, 1994.
19. S. N. Mohammad and H. Morkoc, *Prog. Quant. Electron.* vol. 20, p. 361, 1996.
20. A. S. Barker and M. Ilegems *Phys. Rev. B*, vol. 7, p. 743, 1973.
21. H. P. Maruska and J. J. Tietjen, *Appl. Phys. Lett.* vol. 15, p. 327, 1969.

Resistance of Wires

The following table gives the approximate resistance of various metallic conductors. The values have been computed from the resistivities at 20°C, except as otherwise stated, and for the dimensions of wire indicated. Owing to differences in purity in the case of elements and of composition in alloys, the values can be considered only as approximations.

	Diameter			Diameter	
B. & S. Gauge	mm	mills 1 mil = .001 in	B. & S. gauge	mm	mills 1 mil = .001 in
10	2.588	101.9	26	0.4049	15.94
12	2.053	80.81	27	0.3606	14.20
14	1.628	64.08	28	0.3211	12.64
16	1.291	50.82	30	0.2546	10.03
18	1.024	40.30	32	0.2019	7.950
20	0.8118	31.96	34	0.1601	6.305
22	0.6438	25.35	36	0.1270	5.000
24	0.5106	20.10	40	0.07987	3.145

B. & S. No.	Ohms per cm	Ohms per ft	B. & S. No.	Ohms per cm	Ohms per ft
Advance (0°C) $Q = 48. \times 10^{-6}$ ohm cm			Brass $Q = 7.00 \times 10^{-6}$ ohm cm		
10	.000912	.0278	10	.000133	.00406
12	.00145	.0442	12	.000212	.00645
14	.00231	.0703	14	.000336	.0103
16	.00367	.112	16	.000535	.0163
18	.00583	.178	18	.000850	.0259
20	.00927	.283	20	.00135	.0412
22	.0147	.449	22	.00215	.0655
24	.0234	.715	24	.00342	.104
26	.0373	1.14	26	.00543	.166
27	.0470	1.43	27	.00686	.209
28	.0593	1.81	28	.00864	.263
30	.0942	2.87	30	.0137	.419
32	.150	4.57	32	.0219	.666
34	.238	7.26	34	.0348	1.06
36	.379	11.5	36	.0552	1.68
40	.958	29.2	40	.140	4.26
Aluminum $Q = 2.828 \times 10^{-6}$ ohm cm			Climax $Q = 87. \times 10^{-6}$ ohm cm		
10	.0000538	.00164	10	.00165	.0504
12	.0000855	.00260	12	.00263	.0801
14	.000136	.00414	14	.00418	.127
16	.000216	.00658	16	.00665	.203
18	.000344	.0105	18	.0106	.322
20	.000546	.0167	20	.0168	.512
22	.000869	.0265	22	.0267	.815
24	.00138	.0421	24	.0425	1.30
26	.00220	.0669	26	.0675	2.06
27	.00277	.0844	27	.0852	2.60
28	.00349	.106	28	.107	3.27
30	.00555	.169	30	.171	5.21
32	.00883	.269	32	.272	8.28
34	.0140	.428	34	.432	13.2
36	.0223	.680	36	.687	20.9
40	.0564	1.72	40	1.74	52.9
Constantan (0°C) $Q = 44.1 \times 10^{-6}$ ohm cm			Excello $Q = 92. \times 10^{-6}$ ohm cm		
10	.000838	.0255	10	.00175	.0533
12	.00133	.0406	12	.00278	.0847
14	.00212	.0646	14	.00442	.135
16	.00337	.103	16	.00703	.214
18	.00536	.163	18	.0112	.341
20	.00852	.260	20	.0178	.542
22	.0135	.413	22	.0283	.861
24	.0215	.657	24	.0449	1.37
26	.0342	1.04	26	.0714	2.18
27	.0432	1.32	27	.0901	2.75
28	.0545	1.66	28	.114	3.46
30	.0866	2.64	30	.181	5.51
32	.138	4.20	32	.287	8.75
34	.219	6.67	34	.457	13.9
36	.348	10.6	36	.726	22.1
40	.880	26.8	40	1.84	56.0

B. & S. No.	Ohms per cm	Ohms per ft	B. & S. No.	Ohms per cm	Ohms per ft
Copper, annealed $Q = 1.724 \times 10^{-6}$ ohm cm			German silver $Q = 33. \times 10^{-6}$ ohm cm		
10	.0000328	.000999	10	.000627	.0191
12	.0000521	.00159	12	.000997	.0304
14	.0000828	.00253	14	.00159	.0483
16	.000132	.00401	16	.00252	.0768
18	.000209	.00638	18	.00401	.122
20	.000333	.0102	20	.00638	.194
22	.000530	.0161	22	.0101	.309
24	.000842	.0257	24	.0161	.491
26	.00134	.0408	26	.0256	.781
27	.00169	.0515	27	.0323	.985
28	.00213	.0649	28	.0408	1.24
30	.00339	.103	30	.0648	1.97
32	.00538	.164	32	.103	3.14
34	.00856	.261	34	.164	4.99
36	.0136	.415	36	.260	.794
40	.0344	1.05	40	.659	20.1
Eureka (0°C) $Q = 47. \times 10^{-6}$ ohm cm			Gold $Q = 2.44 \times 10^{-6}$ ohm cm		
10	.000893	.0272	10	.0000464	.00141
12	.00142	.0433	12	.0000737	.00225
14	.00226	0.688	14	.000117	.00357
16	.00359	.109	16	.000186	.00568
18	.00571	.174	18	.000296	.00904
20	.00908	.277	20	.000471	.0144
22	.0144	.440	22	.000750	.0228
24	.0230	.700	24	.00119	.0363
26	.0365	1.11	26	.00189	.0577
27	.0460	1.40	27	.00239	.0728
28	.0580	1.77	28	.00301	.0918
30	.0923	2.81	30	.00479	.146
32	.147	4.47	32	.00762	.232
34	.233	7.11	34	.0121	.369
36	.371	11.3	36	.0193	.587
40	.938	28.6	40	.0487	1.48
Iron $Q = 10. \times 10^{-6}$ ohm cm			Manganin $Q = 44. \times 10^{-6}$ ohm cm		
10	.000190	.00579	10	.000836	.0255
12	.000302	.00921	12	.00133	.0405
14	.000481	.0146	14	.00211	.0644
16	.000764	.0233	16	.00336	.102
18	.00121	.0370	18	.00535	.163
20	.00193	.0589	20	.00850	.259
22	.00307	.0936	22	.0135	.412
24	.00489	.149	24	.0215	.655
26	.00776	.237	26	.0342	1.04
27	.00979	.299	27	.0431	1.31
28	.0123	.376	28	.0543	1.66
30	.0196	.598	30	.0864	2.63
32	.0312	.952	32	.137	4.19
34	.0497	1.51	34	.218	6.66
36	0.789	2.41	36	.347	10.6
40	.200	6.08	40	.878	26.8

B. & S. No.	Ohms per cm	Ohms per ft	B. & S. No.	Ohms per cm	Ohms per ft
Lead $Q = 22. \times 10^{-6}$ ohm cm			Molybdenum $Q = 5.7 \times 10^{-6}$ ohm cm		
10	.000418	.0127	10	.000108	.00330
12	.000665	.0203	12	.000172	.00525
14	.00106	.0322	14	.000274	.00835
16	.00168	.0512	16	.000435	.0133
18	.00267	.0815	18	.000693	.0211
20	.00425	.130	20	.00110	.0336
22	.00676	.206	22	.00175	.0534
24	.0107	.328	24	.00278	.0849
26	.0171	.521	26	.00443	.135
27	.0215	.657	27	.00558	.170
28	.0272	.828	28	.00704	.215
30	.0432	1.32	30	.0112	.341
32	.0687	2.09	32	.0178	.542
34	.109	3.33	34	.0283	.863
36	.174	5.29	36	.0450	1.37
40	.439	13.4	40	.114	3.47
Magnesium $Q = 4.6 \times 10^{-6}$ ohm cm			Monel Metal $Q = 42. \times 10^{-6}$ ohm cm		
10	.0000874	.00267	10	.000798	.0243
12	.000139	.00424	12	.00127	.0387
14	.000221	.00674	14	.00202	.0615
16	.000351	.0107	16	.00321	.0978
18	.000559	.0170	18	.00510	.156
20	.000889	.0271	20	.00811	.247
22	.00141	.0431	22	.0129	.393
24	.00225	.0685	24	.0205	.625
26	.00357	.109	26	.0326	.994
27	.00451	.137	27	.0411	1.25
28	.00568	.173	28	.0519	1.58
30	.00903	.275	30	.0825	2.51
32	.0144	.438	32	.131	4.00
34	.0228	.696	34	.209	6.36
36	.0363	1.11	36	.331	10.1
40	.0918	2.80	40	.838	25.6
*Nichrome $Q = 150. \times 10^{-6}$ ohm cm			Silver (18°C) $Q = 1.629 \times 10^{-6}$ ohm cm		
10	.0021281	.06488	10	.0000310	.000944
12	.0033751	.1029	12	.0000492	.00150
14	.0054054	.1648	14	.0000783	.00239
16	.0085116	.2595	16	.000124	.00379
18	.0138383	.4219	18	.000198	.00603
20	.0216218	.6592	20	.000315	.00959
22	.0346040	1.055	22	.000500	.0153
24	.0548088	1.671	24	.000796	.0243
26	.0875760	2.670	26	.00126	.0386
28	.1394328	4.251	27	.00160	.0486
30	.2214000	6.750	28	.00201	.0613
32	.346040	10.55	30	.00320	.0975
34	.557600	17.00	32	.00509	.155
36	.885600	27.00	34	.00809	.247
38	1.383832	42.19	36	.0129	.392
40	2.303872	70.24	40	.0325	.991

B. & S. No.	Ohms per cm	Ohms per ft	B. & S. No.	Ohms per cm	Ohms per ft
Nickel $Q = 7.8 \times 10^{-6}$ ohm cm			Steel, piano wire (0°C) $Q = 11.8 \times 10^{-6}$ ohm cm		
10	.000148	.00452	10	.000224	.00684
12	.000236	.00718	12	.000357	.0109
14	.000375	.0114	14	.000567	.0173
16	.000596	.0182	16	.000901	.0275
18	.000948	.0289	18	.00143	.0437
20	.00151	.0459	20	.00228	.0695
22	.00240	.0730	22	.00363	.110
24	.00381	.116	24	.00576	.176
26	.00606	.185	26	.00916	.279
27	.00764	.233	27	.0116	.352
28	.00963	.294	28	.0146	.444
30	.0153	.467	30	.0232	.706
32	.0244	.742	32	.0368	1.12
34	.0387	1.18	34	.0586	1.79
36	.0616	1.88	36	.0931	2.84
40	.156	4.75	40	.236	7.18
Platinum $Q = 10. \times 10^{-6}$ ohm cm			Steel, invar (35% Ni) $Q = 81. \times 10^{-6}$ ohm cm		
10	.000190	.00579	10	.00154	.0469
12	.000302	.00921	12	.00245	.0746
14	.000481	.0146	14	.00389	.119
16	.000764	.0233	16	.00619	.189
18	.00121	.0370	18	.00984	.300
20	.00193	.0589	20	.0156	.477
22	.00307	.0936	22	.0249	.758
24	.00489	.149	24	.0396	1.21
26	.00776	.237	26	.0629	1.92
27	.00979	.299	27	.0793	2.42
28	.0123	.376	28	.100	3.05
30	.0196	.598	30	.159	4.85
32	.0312	.952	32	.253	7.71
34	.0497	1.51	34	.402	12.3
36	.0789	2.41	36	.639	19.5
40	.200	6.08	40	1.62	49.3
Tantalum $Q = 15.5 \times 10^{-6}$ ohm cm			Tungsten $Q = 5.51 \times 10^{-6}$ ohm cm		
10	.000295	.00898	10	.000105	.00319
12	.000468	.0143	12	.000167	.00508
14	.000745	.0227	14	.000265	.00807
16	.00118	.0361	16	.000421	.0128
18	.00188	.0574	18	.000669	.0204
20	.00299	.0913	20	.00106	.0324
22	.00476	.145	22	.00169	.0516
24	.00757	.231	24	.00269	.0820
26	.0120	.367	26	.00428	.130
27	.0152	.463	27	.00540	.164
28	.0191	.583	28	.00680	.207
30	.0304	.928	30	.0108	.330
32	.0484	1.47	32	.0172	.524
34	.0770	2.35	34	.0274	.834
36	.122	3.73	36	.0435	1.33
40	.309	9.43	40	.110	3.35

B. & S. No.	Ohms per cm	Ohms per ft	B. & S. No.	Ohms per cm	Ohms per ft
Tin $Q = 11.5 \times 10^{-6}$ ohm cm			Zinc (0°C) $Q = 5.75 \times 10^{-6}$ ohm cm		
10	.000219	.00666	10	.000109	.00333
12	.000348	.0106	12	.000174	.00530
14	.000553	.0168	14	.000276	.00842
16	.000879	.0268	16	.000439	.0134
18	.00140	.0426	18	.000699	.0213
20	.00222	.0677	20	.00111	.0339
22	.00353	.108	22	.00177	.0538
24	.00562	.171	24	.00281	.0856
26	.00893	.272	26	.00446	.136
27	.0113	.343	27	.00563	.172
28	.0142	.433	28	.00710	.216
30	.0226	.688	30	.0113	.344
32	.0359	1.09	32	.0180	.547
34	.0571	1.74	34	.0286	.870
36	.0908	2.77	36	.0454	1.38
40	.230	7.00	40	.115	3.50

Credits

Except for the Properties of Semiconductors section, material in Appendix A was reprinted from the following sources:

D. R. Lide, Ed., *CRC Handbook of Chemistry and Physics,* 76th ed., Boca Raton, Fla.: CRC Press, 1992: International System of Units (SI), conversion constants and multipliers (conversion of temperatures), symbols and terminology for physical and chemical quantities, fundamental physical constants, classification of electromagnetic radiation.

W. H. Beyer, Ed., *CRC Standard Mathematical Tables and Formulae,* 29th ed., Boca Raton, Fla.: CRC Press, 1991: Greek alphabet, conversion constants and multipliers (recommended decimal multiples and submultiples, metric to English, English to metric, general, temperature factors), physical constants, series expansion, integrals, the Fourier transforms, numerical methods, probability, positional notation.

R. J. Tallarida, *Pocket Book of Integrals and Mathematical Formulas,* 2nd ed., Boca Raton, Fla.: CRC Press, 1991: Elementary algebra and geometry; determinants, matrices, and linear systems of equations; trigonometry; analytic geometry; series; differential calculus; integral calculus; vector analysis; special functions; statistics; tables of probability and statistics; table of derivatives.

J. F. Pankow, *Aquatic Chemistry Concepts,* Chelsea, Mich.: Lewis Publishers, 1991: Periodic table of the elements.

J. Shackelford and W. Alexander, Eds., *CRC Materials Science and Engineering Handbook,* Boca Raton, Fla.: CRC Press, 1992: Electrical resistivity of selected alloy cast irons, resistivity of selected ceramics.

Appendix B
Microwave Engineering Appendix

John P. Wendler
Tyco Electronics
Wireless Network Solutions

Attenuator Design Values

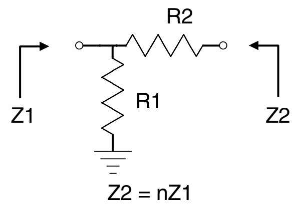


FIGURE B.1 Equivalent circuit for a minimum loss pad.

TABLE 1 Minimum Loss Matching Pad Resistance Values as a Function of Transformation Ratio For Z1 = 1 Ohm, 50 Ohms, and 75 Ohms

n Z2/Z1	r1 Z1 = 1	r2 Z1 = 1	R1 Z1 = 50	R2 Z1 = 50	R1 Z1 = 75	R2 Z1 = 75	Loss [dB]
1.1	3.3166	0.3317	165.8	16.6	248.7	24.9	2.7
1.2	2.4495	0.4899	122.5	24.5	183.7	36.7	3.8
1.3	2.0817	0.6245	104.1	31.2	156.1	46.8	4.5
1.4	1.8708	0.7483	93.5	37.4	140.3	56.1	5.2
1.5	1.7321	0.8660	86.6	43.3	129.9	65.0	5.7
1.6	1.6330	0.9798	81.6	49.0	122.5	73.5	6.2
1.7	1.5584	1.0909	77.9	54.5	116.9	81.8	6.6
1.8	1.5000	1.2000	75.0	60.0	112.5	90.0	7.0
1.9	1.4530	1.3077	72.6	65.4	109.0	98.1	7.3
2.0	1.4142	1.4142	70.7	70.7	106.1	106.1	7.7
2.1	1.3817	1.5199	69.1	76.0	103.6	114.0	8.0
2.2	1.3540	1.6248	67.7	81.2	101.6	121.9	8.2
2.3	1.3301	1.7292	66.5	86.5	99.8	129.7	8.5
2.4	1.3093	1.8330	65.5	91.7	98.2	137.5	8.7
2.5	1.2910	1.9365	64.5	96.8	96.8	145.2	9.0
2.6	1.2748	2.0396	63.7	102.0	95.6	153.0	9.2
2.7	1.2603	2.1424	63.0	107.1	94.5	160.7	9.4
2.8	1.2472	2.2450	62.4	112.2	93.5	168.4	9.6
2.9	1.2354	2.3473	61.8	117.4	92.7	176.1	9.8
3.0	1.2247	2.4495	61.2	122.5	91.9	183.7	10.0
3.1	1.2150	2.5515	60.7	127.6	91.1	191.4	10.1

TABLE 1 (continued)

n Z2/Z1	r1 Z1 = 1	r2 Z1 = 1	R1 Z1 = 50	R2 Z1 = 50	R1 Z1 = 75	R2 Z1 = 75	Loss [dB]
3.2	1.2060	2.6533	60.3	132.7	90.5	199.0	10.3
3.3	1.1978	2.7550	59.9	137.7	89.8	206.6	10.5
3.4	1.1902	2.8566	59.5	142.8	89.3	214.2	10.6
3.5	1.1832	2.9580	59.2	147.9	88.7	221.9	10.8
3.6	1.1767	3.0594	58.8	153.0	88.3	229.5	10.9
3.7	1.1706	3.1607	58.5	158.0	87.8	237.1	11.0
3.8	1.1650	3.2619	58.2	163.1	87.4	244.6	11.2
3.9	1.1597	3.3630	58.0	168.2	87.0	252.2	11.3
4.0	1.1547	3.4641	57.7	173.2	86.6	259.8	11.4
4.1	1.1500	3.5651	57.5	178.3	86.3	267.4	11.6
4.2	1.1456	3.6661	57.3	183.3	85.9	275.0	11.7
4.3	1.1415	3.7670	57.1	188.3	85.6	282.5	11.8
4.4	1.1376	3.8678	56.9	193.4	85.3	290.1	11.9
4.5	1.1339	3.9686	56.7	198.4	85.0	297.6	12.0
4.6	1.1304	4.0694	56.5	203.5	84.8	305.2	12.1
4.7	1.1271	4.1701	56.4	208.5	84.5	312.8	12.2
4.8	1.1239	4.2708	56.2	213.5	84.3	320.3	12.3
4.9	1.1209	4.3715	56.0	218.6	84.1	327.9	12.4
5.0	1.1180	4.4721	55.9	223.6	83.9	335.4	12.5
5.1	1.1153	4.5727	55.8	228.6	83.6	343.0	12.6
5.2	1.1127	4.6733	55.6	233.7	83.5	350.5	12.7
5.3	1.1102	4.7739	55.5	238.7	83.3	358.0	12.8
5.4	1.1078	4.8744	55.4	243.7	83.1	365.6	12.9
5.5	1.1055	4.9749	55.3	248.7	82.9	373.1	13.0
5.6	1.1034	5.0754	55.2	253.8	82.8	380.7	13.1
5.7	1.1013	5.1759	55.1	258.8	82.6	388.2	13.2
5.8	1.0992	5.2764	55.0	263.8	82.4	395.7	13.3
5.9	1.0973	5.3768	54.9	268.8	82.3	403.3	13.3
6.0	1.0954	5.4772	54.8	273.9	82.2	410.8	13.4
6.1	1.0937	5.5776	54.7	278.9	82.0	418.3	13.5
6.2	1.0919	5.6780	54.6	283.9	81.9	425.9	13.6
6.3	1.0903	5.7784	54.5	288.9	81.8	433.4	13.6
6.4	1.0887	5.8788	54.4	293.9	81.6	440.9	13.7
6.5	1.0871	5.9791	54.4	299.0	81.5	448.4	13.8
6.6	1.0856	6.0795	54.3	304.0	81.4	456.0	13.9
6.7	1.0842	6.1798	54.2	309.0	81.3	463.5	13.9
6.8	1.0828	6.2801	54.1	314.0	81.2	471.0	14.0
6.9	1.0814	6.3804	54.1	319.0	81.1	478.5	14.1
7.0	1.0801	6.4807	54.0	324.0	81.0	486.1	14.1
7.1	1.0789	6.5810	53.9	329.1	80.9	493.6	14.2
7.2	1.0776	6.6813	53.9	334.1	80.8	501.1	14.3
7.3	1.0764	6.7816	53.8	339.1	80.7	508.6	14.3
7.4	1.0753	6.8819	53.8	344.1	80.6	516.1	14.4
7.5	1.0742	6.9821	53.7	349.1	80.6	523.7	14.5
7.6	1.0731	7.0824	53.7	354.1	80.5	531.2	14.5
7.7	1.0720	7.1826	53.6	359.1	80.4	538.7	14.6
7.8	1.0710	7.2829	53.6	364.1	80.3	546.2	14.6
7.9	1.0700	7.3831	53.5	369.2	80.3	553.7	14.7
8.0	1.0690	7.4833	53.5	374.2	80.2	561.2	14.8
8.1	1.0681	7.5835	53.4	379.2	80.1	568.8	14.8
8.2	1.0672	7.6837	53.4	384.2	80.0	576.3	14.9
8.3	1.0663	7.7840	53.3	389.2	80.0	583.8	14.9
8.4	1.0654	7.8842	53.3	394.2	79.9	591.3	15.0
8.5	1.0646	7.9844	53.2	399.2	79.8	598.8	15.0
8.6	1.0638	8.0846	53.2	404.2	79.8	606.3	15.1

TABLE 1 (continued)

n Z2/Z1	r1 Z1 = 1	r2 Z1 = 1	R1 Z1 = 50	R2 Z1 = 50	R1 Z1 = 75	R2 Z1 = 75	Loss [dB]
8.7	1.0630	8.1847	53.1	409.2	79.7	613.9	15.2
8.8	1.0622	8.2849	53.1	414.2	79.7	621.4	15.2
8.9	1.0614	8.3851	53.1	419.3	79.6	628.9	15.3
9.0	1.0607	8.4853	53.0	424.3	79.5	636.4	15.3
9.1	1.0599	8.5855	53.0	429.3	79.5	643.9	15.4
9.2	1.0592	8.6856	53.0	434.3	79.4	651.4	15.4
9.3	1.0585	8.7858	52.9	439.3	79.4	658.9	15.5
9.4	1.0579	8.8859	52.9	444.3	79.3	666.4	15.5
9.5	1.0572	8.9861	52.9	449.3	79.3	674.0	15.6
9.6	1.0565	9.0863	52.8	454.3	79.2	681.5	15.6
9.7	1.0559	9.1864	52.8	459.3	79.2	689.0	15.7
9.8	1.0553	9.2865	52.8	464.3	79.1	696.5	15.7
9.9	1.0547	9.3867	52.7	469.3	79.1	704.0	15.7
10.0	1.0541	9.4868	52.7	474.3	79.1	711.5	15.8

$$R_1 = \frac{Z_1(n+\sqrt{n^2-n})}{n-1+\sqrt{n^2-n}} \qquad R_2 = Z_1\sqrt{n^2-n} \qquad \frac{P_O}{P_A} = \frac{1}{n}\left(\frac{1}{1+\sqrt{\frac{1}{n}}}\right)^2$$

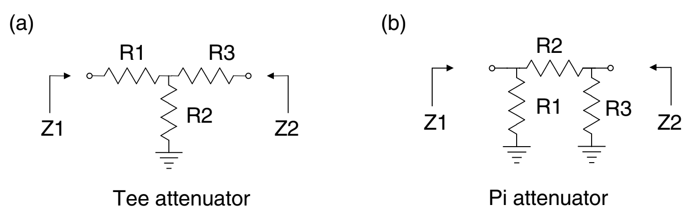


FIGURE B.2 (a) Equivalent circuit for a Tee attenuator; (b) Equivalent circuit for a Pi attenuator.

TABLE 2 Tee- and Pi-Pad Resistor Values for Zo = 1 Ohm and Zo = 50 Ohms

Loss [dB]	Voltage Atten	r1, r3, g1, g3 Z1 = Z2 = 1	r2, g2 Z1 = Z2 = 1	Tee R1, R3 Z1 = Z2 = 50	Tee R2 Z1 = Z2 = 50	Pi R1, R3 Z1 = Z2 = 50	Pi R2 Z1 = Z2 = 50
0.1	0.98855	0.0058	86.8570	0.3	4342.8	8686.0	0.6
0.2	0.97724	0.0115	43.4256	0.6	2171.3	4343.1	1.2
0.3	0.96605	0.0173	28.9472	0.9	1447.4	2895.6	1.7
0.4	0.95499	0.0230	21.7071	1.2	1085.4	2171.9	2.3
0.5	0.94406	0.0288	17.3622	1.4	868.1	1737.7	2.9
0.6	0.93325	0.0345	14.4650	1.7	723.2	1448.2	3.5
0.7	0.92257	0.0403	12.3950	2.0	619.7	1241.5	4.0
0.8	0.91201	0.0460	10.8420	2.3	542.1	1086.5	4.6
0.9	0.90157	0.0518	9.6337	2.6	481.7	966.0	5.2
1.0	0.89125	0.0575	8.6667	2.9	433.3	869.5	5.8
1.2	0.87096	0.0690	7.2153	3.4	360.8	725.0	6.9
1.4	0.85114	0.0804	6.1774	4.0	308.9	621.8	8.1
1.6	0.83176	0.0918	5.3981	4.6	269.9	544.4	9.3
1.8	0.81283	0.1032	4.7911	5.2	239.6	484.3	10.4
2.0	0.79433	0.1146	4.3048	5.7	215.2	436.2	11.6

TABLE 2 (continued)

Loss [dB]	Voltage Atten	r1, r3, g1, g3 Z1 = Z2 = 1	r2, g2 Z1 = Z2 = 1	Tee R1, R3 Z1 = Z2 = 50	Tee R2 Z1 = Z2 = 50	Pi R1, R3 Z1 = Z2 = 50	Pi R2 Z1 = Z2 = 50
2.2	0.77625	0.1260	3.9062	6.3	195.3	396.9	12.8
2.4	0.75858	0.1373	3.5735	6.9	178.7	364.2	14.0
2.6	0.74131	0.1486	3.2914	7.4	164.6	336.6	15.2
2.8	0.72444	0.1598	3.0490	8.0	152.5	312.9	16.4
3.0	0.70795	0.1710	2.8385	8.5	141.9	292.4	17.6
3.2	0.69183	0.1822	2.6539	9.1	132.7	274.5	18.8
3.4	0.67608	0.1933	2.4906	9.7	124.5	258.7	20.1
3.6	0.66069	0.2043	2.3450	10.2	117.3	244.7	21.3
3.8	0.64565	0.2153	2.2144	10.8	110.7	232.2	22.6
4.0	0.63096	0.2263	2.0966	11.3	104.8	221.0	23.8
4.2	0.61660	0.2372	1.9896	11.9	99.5	210.8	25.1
4.4	0.60256	0.2480	1.8921	12.4	94.6	201.6	26.4
4.6	0.58884	0.2588	1.8028	12.9	90.1	193.2	27.7
4.8	0.57544	0.2695	1.7206	13.5	86.0	185.5	29.1
5.0	0.56234	0.2801	1.6448	14.0	82.2	178.5	30.4
5.5	0.53088	0.3064	1.4785	15.3	73.9	163.2	33.8
6.0	0.50119	0.3323	1.3386	16.6	66.9	150.5	37.4
6.5	0.47315	0.3576	1.2193	17.9	61.0	139.8	41.0
7.0	0.44668	0.3825	1.1160	19.1	55.8	130.7	44.8
7.5	0.42170	0.4068	1.0258	20.3	51.3	122.9	48.7
8.0	0.39811	0.4305	0.9462	21.5	47.3	116.1	52.8
8.5	0.37584	0.4537	0.8753	22.7	43.8	110.2	57.1
9.0	0.35481	0.4762	0.8118	23.8	40.6	105.0	61.6
9.5	0.33497	0.4982	0.7546	24.9	37.7	100.4	66.3
10.0	0.31623	0.5195	0.7027	26.0	35.1	96.2	71.2
10.5	0.29854	0.5402	0.6555	27.0	32.8	92.6	76.3
11.0	0.28184	0.5603	0.6123	28.0	30.6	89.2	81.7
11.5	0.26607	0.5797	0.5727	29.0	28.6	86.3	87.3
12.0	0.25119	0.5985	0.5362	29.9	26.8	83.5	93.2
12.5	0.23714	0.6166	0.5025	30.8	25.1	81.1	99.5
13.0	0.22387	0.6342	0.4714	31.7	23.6	78.8	106.1
13.5	0.21135	0.6511	0.4425	32.6	22.1	76.8	113.0
14.0	0.19953	0.6673	0.4156	33.4	20.8	74.9	120.3
14.5	0.18836	0.6830	0.3906	34.1	19.5	73.2	128.0
15.0	0.17783	0.6980	0.3673	34.9	18.4	71.6	136.1
15.5	0.16788	0.7125	0.3455	35.6	17.3	70.2	144.7
16.0	0.15849	0.7264	0.3251	36.3	16.3	68.8	153.8
16.5	0.14962	0.7397	0.3061	37.0	15.3	67.6	163.3
17.0	0.14125	0.7525	0.2883	37.6	14.4	66.4	173.5
17.5	0.13335	0.7647	0.2715	38.2	13.6	65.4	184.1
18.0	0.12589	0.7764	0.2558	38.8	12.8	64.4	195.4
18.5	0.11885	0.7875	0.2411	39.4	12.1	63.5	207.4
19.0	0.11220	0.7982	0.2273	39.9	11.4	62.6	220.0
19.5	0.10593	0.8084	0.2143	40.4	10.7	61.8	233.4
20.0	0.10000	0.8182	0.2020	40.9	10.1	61.1	247.5
20.5	0.09441	0.8275	0.1905	41.4	9.5	60.4	262.5
21.0	0.08913	0.8363	0.1797	41.8	9.0	59.8	278.3
21.5	0.08414	0.8448	0.1695	42.2	8.5	59.2	295.0
22.0	0.07943	0.8528	0.1599	42.6	8.0	58.6	312.7
22.5	0.07499	0.8605	0.1508	43.0	7.5	58.1	331.5
23.0	0.07079	0.8678	0.1423	43.4	7.1	57.6	351.4
23.5	0.06683	0.8747	0.1343	43.7	6.7	57.2	372.4
24.0	0.06310	0.8813	0.1267	44.1	6.3	56.7	394.6
24.5	0.05957	0.8876	0.1196	44.4	6.0	56.3	418.2
25.0	0.05623	0.8935	0.1128	44.7	5.6	56.0	443.2

TABLE 2 (continued)

Loss [dB]	Voltage Atten	r1, r3, g1, g3 Z1 = Z2 = 1	r2, g2 Z1 = Z2 = 1	Tee R1, R3 Z1 = Z2 = 50	Tee R2 Z1 = Z2 = 50	Pi R1, R3 Z1 = Z2 = 50	Pi R2 Z1 = Z2 = 50
26.0	0.05012	0.9045	0.1005	45.2	5.0	55.3	497.6
27.0	0.04467	0.9145	0.0895	45.7	4.5	54.7	558.6
28.0	0.03981	0.9234	0.0797	46.2	4.0	54.1	627.0
29.0	0.03548	0.9315	0.0711	46.6	3.6	53.7	703.7
30.0	0.03162	0.9387	0.0633	46.9	3.2	53.3	789.8
31.0	0.02818	0.9452	0.0564	47.3	2.8	52.9	886.3
32.0	0.02512	0.9510	0.0503	47.5	2.5	52.6	994.6
33.0	0.02239	0.9562	0.0448	47.8	2.2	52.3	1116.1
34.0	0.01995	0.9609	0.0399	48.0	2.0	52.0	1252.5
35.0	0.01778	0.9651	0.0356	48.3	1.8	51.8	1405.4
36.0	0.01585	0.9688	0.0317	48.4	1.6	51.6	1577.0
37.0	0.01413	0.9721	0.0283	48.6	1.4	51.4	1769.5
38.0	0.01259	0.9751	0.0252	48.8	1.3	51.3	1985.5
39.0	0.01122	0.9778	0.0224	48.9	1.1	51.1	2227.8
40.0	0.01000	0.9802	0.0200	49.0	1.0	51.0	2499.8
41.0	0.00891	0.9823	0.0178	49.1	0.9	50.9	2804.8
42.0	0.00794	0.9842	0.0159	49.2	0.8	50.8	3147.1
43.0	0.00708	0.9859	0.0142	49.3	0.7	50.7	3531.2
44.0	0.00631	0.9875	0.0126	49.4	0.6	50.6	3962.1
45.0	0.00562	0.9888	0.0112	49.4	0.6	50.6	4445.6

Note: P_i values are duals of Tee values.

$$a = \sqrt{\frac{P_{z_2}}{P_{z_1}}} \qquad R_{1T} = \left(\frac{2}{(1-a^2)} - 1\right) Z_1 - \frac{2a}{(1-a^2)}\sqrt{Z_1 Z_2}$$

$$R_{2T} = 2\sqrt{Z_1 Z_2}\frac{a}{(1-a^2)}$$

$$R_{3T} = \left(\frac{2}{(1-a^2)} - 1\right) Z_1 - \frac{2a}{(1-a^2)}\sqrt{Z_1 Z_2}$$

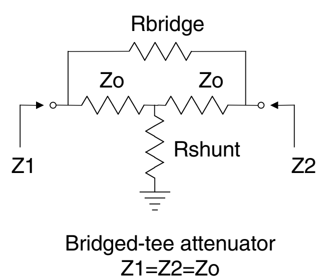


FIGURE B.3 Equivalent circuit for a Bridged-T attenuator.

TABLE 3 Bridged-T Attenuator Resistance Values for Zo = 1 Ohm, 50 Ohms, 75 Ohms

Loss [dB]	Voltage Atten	Bridge Arm Z1 = Z2 = 1	Shunt Arm Z1 = Z2 = 1	Bridge Arm Z1 = Z2 = 50	Shunt Arm Z1 = Z2 = 50	Bridge Arm Z1 = Z2 = 75	Shunt Arm Z1 = Z2 = 75
0.1	0.98855	0.0116	86.3599	0.6	4318.0	6477.0	0.9
0.2	0.97724	0.0233	42.9314	1.2	2146.6	3219.9	1.7
0.3	0.96605	0.0351	28.4558	1.8	1422.8	2134.2	2.6
0.4	0.95499	0.0471	21.2186	2.4	1060.9	1591.4	3.5
0.5	0.94406	0.0593	16.8766	3.0	843.8	1265.7	4.4
0.6	0.93325	0.0715	13.9822	3.6	699.1	1048.7	5.4
0.7	0.92257	0.0839	11.9151	4.2	595.8	893.6	6.3

TABLE 3 (continued)

Loss [dB]	Voltage Atten	Bridge Arm Z1 = Z2 = 1	Shunt Arm Z1 = Z2 = 1	Bridge Arm Z1 = Z2 = 50	Shunt Arm Z1 = Z2 = 50	Bridge Arm Z1 = Z2 = 75	Shunt Arm Z1 = Z2 = 75
0.8	0.91201	0.0965	10.3650	4.8	518.3	777.4	7.2
0.9	0.90157	0.1092	9.1596	5.5	458.0	687.0	8.2
1.0	0.89125	0.1220	8.1955	6.1	409.8	614.7	9.2
1.2	0.87096	0.1482	6.7498	7.4	337.5	506.2	11.1
1.4	0.85114	0.1749	5.7176	8.7	285.9	428.8	13.1
1.6	0.83176	0.2023	4.9440	10.1	247.2	370.8	15.2
1.8	0.81283	0.2303	4.3428	11.5	217.1	325.7	17.3
2.0	0.79433	0.2589	3.8621	12.9	193.1	289.7	19.4
2.2	0.77625	0.2882	3.4692	14.4	173.5	260.2	21.6
2.4	0.75858	0.3183	3.1421	15.9	157.1	235.7	23.9
2.6	0.74131	0.3490	2.8656	17.4	143.3	214.9	26.2
2.8	0.72444	0.3804	2.6289	19.0	131.4	197.2	28.5
3.0	0.70795	0.4125	2.4240	20.6	121.2	181.8	30.9
3.2	0.69183	0.4454	2.2450	22.3	112.2	168.4	33.4
3.4	0.67608	0.4791	2.0872	24.0	104.4	156.5	35.9
3.6	0.66069	0.5136	1.9472	25.7	97.4	146.0	38.5
3.8	0.64565	0.5488	1.8221	27.4	91.1	136.7	41.2
4.0	0.63096	0.5849	1.7097	29.2	85.5	128.2	43.9
4.2	0.61660	0.6218	1.6082	31.1	80.4	120.6	46.6
4.4	0.60256	0.6596	1.5161	33.0	75.8	113.7	49.5
4.6	0.58884	0.6982	1.4322	34.9	71.6	107.4	52.4
4.8	0.57544	0.7378	1.3554	36.9	67.8	101.7	55.3
5.0	0.56234	0.7783	1.2849	38.9	64.2	96.4	58.4
5.5	0.53088	0.8836	1.1317	44.2	56.6	84.9	66.3
6.0	0.50119	0.9953	1.0048	49.8	50.2	75.4	74.6
6.5	0.47315	1.1135	0.8981	55.7	44.9	67.4	83.5
7.0	0.44668	1.2387	0.8073	61.9	40.4	60.5	92.9
7.5	0.42170	1.3714	0.7292	68.6	36.5	54.7	102.9
8.0	0.39811	1.5119	0.6614	75.6	33.1	49.6	113.4
8.5	0.37584	1.6607	0.6021	83.0	30.1	45.2	124.6
9.0	0.35481	1.8184	0.5499	90.9	27.5	41.2	136.4
9.5	0.33497	1.9854	0.5037	99.3	25.2	37.8	148.9
10.0	0.31623	2.1623	0.4625	108.1	23.1	34.7	162.2
10.5	0.29854	2.3497	0.4256	117.5	21.3	31.9	176.2
11.0	0.28184	2.5481	0.3924	127.4	19.6	29.4	191.1
11.5	0.26607	2.7584	0.3625	137.9	18.1	27.2	206.9
12.0	0.25119	2.9811	0.3354	149.1	16.8	25.2	223.6
12.5	0.23714	3.2170	0.3109	160.8	15.5	23.3	241.3
13.0	0.22387	3.4668	0.2884	173.3	14.4	21.6	260.0
13.5	0.21135	3.7315	0.2680	186.6	13.4	20.1	279.9
14.0	0.19953	4.0119	0.2493	200.6	12.5	18.7	300.9
14.5	0.18836	4.3088	0.2321	215.4	11.6	17.4	323.2
15.0	0.17783	4.6234	0.2163	231.2	10.8	16.2	346.8
15.5	0.16788	4.9566	0.2018	247.8	10.1	15.1	371.7
16.0	0.15849	5.3096	0.1883	265.5	9.4	14.1	398.2
16.5	0.14962	5.6834	0.1759	284.2	8.8	13.2	426.3
17.0	0.14125	6.0795	0.1645	304.0	8.2	12.3	456.0
17.5	0.13335	6.4989	0.1539	324.9	7.7	11.5	487.4
18.0	0.12589	6.9433	0.1440	347.2	7.2	10.8	520.7
18.5	0.11885	7.4140	0.1349	370.7	6.7	10.1	556.0
19.0	0.11220	7.9125	0.1264	395.6	6.3	9.5	593.4
19.5	0.10593	8.4406	0.1185	422.0	5.9	8.9	633.0
20.0	0.10000	9.0000	0.1111	450.0	5.6	8.3	675.0
20.5	0.09441	9.5925	0.1042	479.6	5.2	7.8	719.4
21.0	0.08913	10.2202	0.0978	511.0	4.9	7.3	766.5
21.5	0.08414	10.8850	0.0919	544.3	4.6	6.9	816.4

TABLE 3 (continued)

Loss [dB]	Voltage Atten	Bridge Arm Z1 = Z2 = 1	Shunt Arm Z1 = Z2 = 1	Bridge Arm Z1 = Z2 = 50	Shunt Arm Z1 = Z2 = 50	Bridge Arm Z1 = Z2 = 75	Shunt Arm Z1 = Z2 = 75
22.0	0.07943	11.5893	0.0863	579.5	4.3	6.5	869.2
22.5	0.07499	12.3352	0.0811	616.8	4.1	6.1	925.1
23.0	0.07079	13.1254	0.0762	656.3	3.8	5.7	984.4
23.5	0.06683	13.9624	0.0716	698.1	3.6	5.4	1047.2
24.0	0.06310	14.8489	0.0673	742.4	3.4	5.1	1113.7
24.5	0.05957	15.7880	0.0633	789.4	3.2	4.8	1184.1
25.0	0.05623	16.7828	0.0596	839.1	3.0	4.5	1258.7
26.0	0.05012	18.9526	0.0528	947.6	2.6	4.0	1421.4
27.0	0.04467	21.3872	0.0468	1069.4	2.3	3.5	1604.0
28.0	0.03981	24.1189	0.0415	1205.9	2.1	3.1	1808.9
29.0	0.03548	27.1838	0.0368	1359.2	1.8	2.8	2038.8
30.0	0.03162	30.6228	0.0327	1531.1	1.6	2.4	2296.7
31.0	0.02818	34.4813	0.0290	1724.1	1.5	2.2	2586.1
32.0	0.02512	38.8107	0.0258	1940.5	1.3	1.9	2910.8
33.0	0.02239	43.6684	0.0229	2183.4	1.1	1.7	3275.1
34.0	0.01995	49.1187	0.0204	2455.9	1.0	1.5	3683.9
35.0	0.01778	55.2341	0.0181	2761.7	0.9	1.4	4142.6
36.0	0.01585	62.0957	0.0161	3104.8	0.8	1.2	4657.2
37.0	0.01413	69.7946	0.0143	3489.7	0.7	1.1	5234.6
38.0	0.01259	78.4328	0.0127	3921.6	0.6	1.0	5882.5
39.0	0.01122	88.1251	0.0113	4406.3	0.6	0.9	6609.4
40.0	0.01000	99.0000	0.0101	4950.0	0.5	0.8	7425.0
41.0	0.00891	111.2018	0.0090	5560.1	0.4	0.7	8340.1
42.0	0.00794	124.8925	0.0080	6244.6	0.4	0.6	9366.9
43.0	0.00708	140.2538	0.0071	7012.7	0.4	0.5	10519.0
44.0	0.00631	157.4893	0.0063	7874.5	0.3	0.5	11811.7
45.0	0.00562	176.8279	0.0057	8841.4	0.3	0.4	13262.1

Return Loss, Reflection Coefficient, VSWR, and Mismatch Loss

TABLE 4 Conversion Between Return Loss, Reflection Coefficient, VSWR, and Mismatch Loss

Return Loss [dB]	Reflection Coefficient (Rho)	VSWR ( ):1	Mismatch Loss [dB]	Return Loss [dB]	Reflection Coefficient (Rho)	VSWR ( ):1	Mismatch Loss [dB]
Infinite	0.0000	1.00	0.00	33.00	0.0224	1.05	0.00
50.00	0.0032	1.01	0.00	32.00	0.0251	1.05	0.00
49.00	0.0035	1.01	0.00	31.00	0.0282	1.06	0.00
48.00	0.0040	1.01	0.00	30.00	0.0316	1.07	0.00
47.00	0.0045	1.01	0.00	29.00	0.0355	1.07	0.01
46.00	0.0050	1.01	0.00	28.00	0.0398	1.08	0.01
45.00	0.0056	1.01	0.00	27.00	0.0447	1.09	0.01
44.00	0.0063	1.01	0.00	26.00	0.0501	1.11	0.01
43.00	0.0071	1.01	0.00	25.00	0.0562	1.12	0.01
42.00	0.0079	1.02	0.00	24.00	0.0631	1.13	0.02
41.00	0.0089	1.02	0.00	23.00	0.0708	1.15	0.02
40.00	0.0100	1.02	0.00	22.00	0.0794	1.17	0.03
39.00	0.0112	1.02	0.00	21.00	0.0891	1.20	0.03
38.00	0.0126	1.03	0.00	20.00	0.1000	1.22	0.04
37.00	0.0141	1.03	0.00	19.50	0.1059	1.24	0.05
36.00	0.0158	1.03	0.00	19.00	0.1122	1.25	0.06
35.00	0.0178	1.04	0.00	18.50	0.1189	1.27	0.06
34.00	0.0200	1.04	0.00	18.00	0.1259	1.29	0.07
17.50	0.1334	1.31	0.08	6.02	0.5000	3.00	1.25
17.00	0.1413	1.33	0.09	6.00	0.5012	3.01	1.26
16.50	0.1496	1.35	0.10	5.80	0.5129	3.11	1.33
16.00	0.1585	1.38	0.11	5.60	0.5248	3.21	1.40
15.50	0.1679	1.40	0.12	5.40	0.5370	3.32	1.48
15.00	0.1778	1.43	0.14	5.20	0.5495	3.44	1.56
14.50	0.1884	1.46	0.16	5.11	0.5556	3.50	1.60
14.00	0.1995	1.50	0.18	5.00	0.5623	3.57	1.65
13.50	0.2113	1.54	0.20	4.80	0.5754	3.71	1.75
13.00	0.2239	1.58	0.22	4.60	0.5888	3.86	1.85
12.50	0.2371	1.62	0.25	4.44	0.6000	4.00	1.94
12.00	0.2512	1.67	0.28	4.40	0.6026	4.03	1.96
11.50	0.2661	1.73	0.32	4.20	0.6166	4.22	2.08
11.00	0.2818	1.78	0.36	4.00	0.6310	4.42	2.20
10.50	0.2985	1.85	0.41	3.93	0.6364	4.50	2.25
10.00	0.3162	1.92	0.46	3.80	0.6457	4.64	2.34
9.80	0.3236	1.96	0.48	3.60	0.6607	4.89	2.49
9.60	0.3311	1.99	0.50	3.52	0.6667	5.00	2.55
9.54	0.3333	2.00	0.51	3.40	0.6761	5.17	2.65
9.40	0.3388	2.03	0.53	3.20	0.6918	5.49	2.83
9.20	0.3467	2.06	0.56	3.00	0.7079	5.85	3.02
9.00	0.3548	2.10	0.58	2.80	0.7244	6.26	3.23
8.80	0.3631	2.14	0.61	2.60	0.7413	6.73	3.46
8.60	0.3715	2.18	0.65	2.40	0.7586	7.28	3.72
8.40	0.3802	2.23	0.68	2.20	0.7762	7.94	4.01
8.20	0.3890	2.27	0.71	2.00	0.7943	8.72	4.33
8.00	0.3981	2.32	0.75	1.80	0.8128	9.69	4.69
7.80	0.4074	2.37	0.79	1.74	0.8182	10.00	4.81
7.60	0.4169	2.43	0.83	1.60	0.8318	10.89	5.11
7.40	0.4266	2.49	0.87	1.40	0.8511	12.44	5.60
7.36	0.4286	2.50	0.88	1.20	0.8710	14.50	6.17
7.20	0.4365	2.55	0.92	1.00	0.8913	17.39	6.87
7.00	0.4467	2.61	0.97	0.80	0.9120	21.73	7.74
6.80	0.4571	2.68	1.02	0.60	0.9333	28.96	8.89
6.60	0.4677	2.76	1.07	0.40	0.9550	43.44	10.56
6.40	0.4786	2.84	1.13	0.20	0.9772	86.86	13.47
6.20	0.4898	2.92	1.19	0.00	1.0000	Infinite	Infinite

Notes:
1. Return Loss = -20*log(|Rho|)
2. Mismatch Loss = -10*log(1-|Rho|^2)
3. VSWR = (1+|Rho|)/(1-|Rho|)

Waveguide Components

TABLE 5 Waveguide Performance and Dimensions

EIA WR-	Mil-W-85E RG( )/U	TE10 Cutoff Frequency [GHz]	Recommended Frequency Range Min [GHz]	Recommended Frequency Range Max [GHz]	Theoretical Attenuation Fmin [dB/100']	Theoretical Attenuation Fmax [dB/100']	Inside Dimensions a [inches]	Inside Dimensions b [Inches]	Tolerance ± [Inches]	Outside Dimensions [Inches]	Outside Dimensions [Inches]	Tolerance ± [Inches]	Wall Thickness [Inches]	Material	Contact Flange	Choke Flange	Cover Flange	Hole Pattern Figure
3	139	173.5726	220.00	325.00	503.90	352.59	0.0340	0.0170	0.00020	4.156	(Diameter)	0.001		Silver				
4	137	137.2434	170.00	260.00	371.25	246.94	0.0430	0.0215	0.00020	3.156	(Diameter)	0.001		Silver				
5	135	115.7151	140.00	220.00	303.47	190.96	0.0510	0.0255	0.00025	2.156	(Diameter)	0.001		Silver				
7	136	90.7918	110.00	170.00	210.19	133.39	0.0650	0.0325	0.00025	1.156	(Diameter)	0.001		Silver				
8	138	73.7683	90.00	140.00	151.38	97.26	0.080	0.040	0.0003	0.156	(Diameter)	0.001		Silver				
10		59.0147	75.00	110.00			0.100	0.050	0.0005	0.180	0.130	0.002	0.040					1
12		48.3727	60.00	90.00	122.71	82.37	0.122	0.061	0.0005	0.202	0.141	0.002	0.040	Brass			387	1
12	99	48.3727	60.00	90.00	77.46	51.99	0.122	0.061	0.0005	0.202	0.141	0.002	0.040	Silver				1
15		39.8748	50.00	75.00	89.78	61.41	0.148	0.074	0.0010	0.228	0.154	0.002	0.040	Brass			385	1
15	98	39.8748	50.00	75.00	56.67	38.76	0.148	0.074	0.0010	0.228	0.154	0.002	0.040	Silver				1
19		31.3908	40.00	60.00			0.188	0.094	0.0010	0.268	0.174	0.002	0.040					1
22		26.3458	33.00	50.00	48.33	32.89	0.224	0.112	0.0010	0.304	0.192	0.002	0.040	Brass			383	1
22	97	26.3458	33.00	50.00	30.50	20.76	0.224	0.112	0.0010	0.304	0.192	0.002	0.040	Silver				1
28		21.0767	26.50	40.00	28.00	19.20	0.280	0.140	0.0015	0.360	0.220	0.002	0.040	Aluminum				1,2
28		21.0767	26.50	40.00	34.32	23.53	0.280	0.140	0.0015	0.360	0.220	0.002	0.040	Brass		600	599	1,2
28	96	21.0767	26.50	40.00	21.66	14.85	0.280	0.140	0.0015	0.360	0.220	0.002	0.040	Silver				1,2
34		17.3573	22.00	33.00			0.340	0.170	0.0020	0.420	0.250	0.003	0.040					
42	121	14.0511	18.00	26.50	16.86	12.40	0.420	0.170	0.0020	0.500	0.250	0.003	0.040	Aluminum	425	598	597	1,2
42	53	14.0511	18.00	26.50	20.66	15.19	0.420	0.170	0.0020	0.500	0.250	0.003	0.040	Brass	425	596	595	1,2
42	66	14.0511	18.00	26.50	13.04	9.59	0.420	0.170	0.0020	0.500	0.250	0.003	0.040	Silver	425	596A	595	1,2
51		11.5715	15.00	22.00			0.510	0.255	0.0025	0.590	0.335	0.003	0.040					2
62		9.4879	12.40	18.00	7.88	5.80	0.622	0.311	0.0025	0.702	0.391	0.003	0.040	Aluminum				2
62	91	9.4879	12.40	18.00	9.66	7.11	0.622	0.311	0.0025	0.702	0.391	0.003	0.040	Brass		541	419	2
62	107	9.4879	12.40	18.00	6.10	4.49	0.622	0.311	0.0025	0.702	0.391	0.003	0.040	Silver				2
75		7.8686	10.00	15.00			0.750	0.375	0.003	0.850	0.475	0.003	0.050					2
90	67	6.5572	8.20	12.40	5.29	3.66	0.900	0.400	0.003	1.000	0.500	0.003	0.050	Aluminum		136A	135	2
90	52	6.5572	8.20	12.40	6.48	4.49	0.900	0.400	0.003	1.000	0.500	0.003	0.050	Brass		40A	39	2
112	68	5.2598	7.05	10.00	3.39	2.63	1.122	0.497	0.004	1.250	0.625	0.004	0.064	Aluminum		137A	138	2
112	51	5.2598	7.05	10.00	4.15	3.23	1.122	0.497	0.004	1.250	0.625	0.004	0.064	Brass		52A	51	2
137	106	4.3014	5.85	8.20	2.42	1.91	1.372	0.622	0.004	1.500	0.750	0.004	0.064	Aluminum		440A	441	
137	50	4.3014	5.85	8.20	2.96	2.34	1.372	0.622	0.004	1.500	0.750	0.004	0.064	Brass		343A	344	
159		3.7116	4.90	7.05			1.590	0.795	0.004	1.718	0.923	0.004	0.064					

TABLE 5 (continued) Waveguide Performance and Dimensions

EIA WR-	Mil-W-85E RG( )/U	TE10 Cutoff Frequency [GHz]	Recommended Frequency Range Min [GHz]	Recommended Frequency Range Max [GHz]	Theoretical Attenuation Fmin [dB/100']	Theoretical Attenuation Fmax [dB/100']	Inside Dimensions a [inches]	Inside Dimensions b [Inches]	Tolerance ± [Inches]	Outside Dimensions [Inches]	Outside Dimensions [Inches]	Tolerance ± [Inches]	Wall Thickness [Inches]	Material	Contact Flange	Choke Flange	Cover Flange	Hole Pattern Figure
187	95	3.1525	3.95	5.85	1.70	1.18	1.872	0.872	0.005	2.000	1.000	0.005	0.064	Aluminum		406A	407	
187	49	3.1525	3.95	5.85	2.09	1.45	1.872	0.872	0.005	2.000	1.000	0.005	0.064	Brass		148B	149A	
229		2.5771	3.30	4.90			2.290	1.145	0.005	2.418	1.273	0.005	0.064					
284	75	2.0780	2.60	3.95	0.91	0.62	2.840	1.340	0.005	3.000	1.500	0.005	0.080	Aluminum		585	584	
284	48	2.0780	2.60	3.95	1.11	0.76	2.840	1.340	0.005	3.000	1.500	0.005	0.080	Brass		54B	53	
340	113	1.7357	2.20	3.30	0.65	0.45	3.400	1.700	0.005	3.560	1.860	0.005	0.080	Aluminum			554	
340	112	1.7357	2.20	3.30	0.80	0.56	3.400	1.700	0.005	3.560	1.860	0.005	0.080	Brass			553	
430	105	1.3724	1.70	2.60	0.48	0.32	4.300	2.150	0.005	4.460	2.310	0.005	0.080	Aluminum			437A	
430	104	1.3724	1.70	2.60	0.59	0.39	4.300	2.150	0.005	4.460	2.310	0.005	0.080	Brass			435A	
510		1.1572	1.45	2.20			5.100	2.550	0.005	5.260	2.710	0.005	0.080					
650	103	0.9079	1.12	1.70	0.26	0.17	6.500	3.250	0.005	6.660	3.410	0.005	0.080	Aluminum			418A	
650	69	0.9079	1.12	1.70	0.32	0.21	6.500	3.250	0.005	6.660	3.410	0.005	0.080	Brass			417A	
770	205	0.7664	0.96	1.45	0.20	0.13	7.700	3.850	0.005	7.950	4.100	0.005	0.125	Aluminum				
975	204	0.6053	0.75	1.12	0.14	0.09	9.750	4.875	0.010	10.000	5.125	0.010	0.125	Aluminum				
1150	203	0.5132	0.64	0.96	0.11	0.07	11.500	5.750	0.015	11.750	6.000	0.015	0.125	Aluminum				
1500	202	0.3934	0.49	0.75	0.07	0.05	15.000	7.500	0.015	15.250	7.750	0.015	0.125	Aluminum				
1800	201	0.3279	0.41	0.63	0.05	0.04	18.000	9.000	0.020	18.250	9.250	0.020	0.125	Aluminum				
2100		0.2810	0.35	0.53	0.04	0.03	21.000	10.500	0.020	21.250	10.750	0.020	0.125	Aluminum				
2300		0.2566	0.32	0.49	0.04	0.03	23.000	11.500	0.020	23.250	11.750	0.020	0.125	Aluminum				

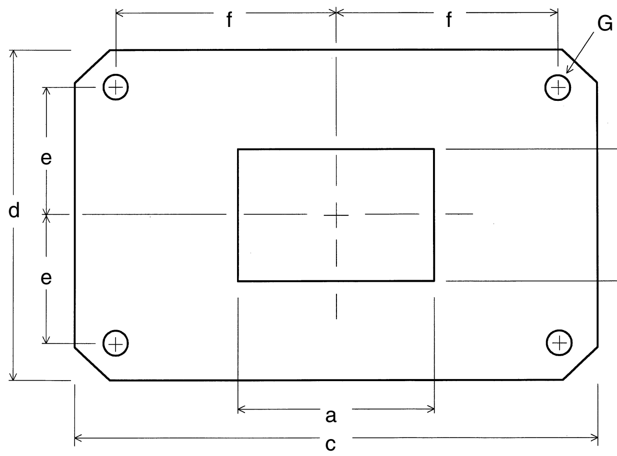
Notes:

1. Conductivity of 63.0e6 Mhos/Meter used for Silver.
2. Conductivity of 37.7e6 Mhos/Meter used for Aluminum.
3. Conductivity of 25.1e6 Mhos/Meter used for Brass.
4. Loss is inversely proportional to the square root of conductivity.

Source:

1. Balanis, C.A., *Advanced Engineering Electromagnetics*, John Wiley & Sons, New York, 1990.
2. Catalog, Microwave Development Labs, Natick, MA.
3. Catalog, Aerowave Inc, Medford, MA.
4. Catalog, Formcraft Tool Co, Chicago, IL.
5. Catalog, Penn Engineering Components, No. Hollywood, CA.

(a)



(b)

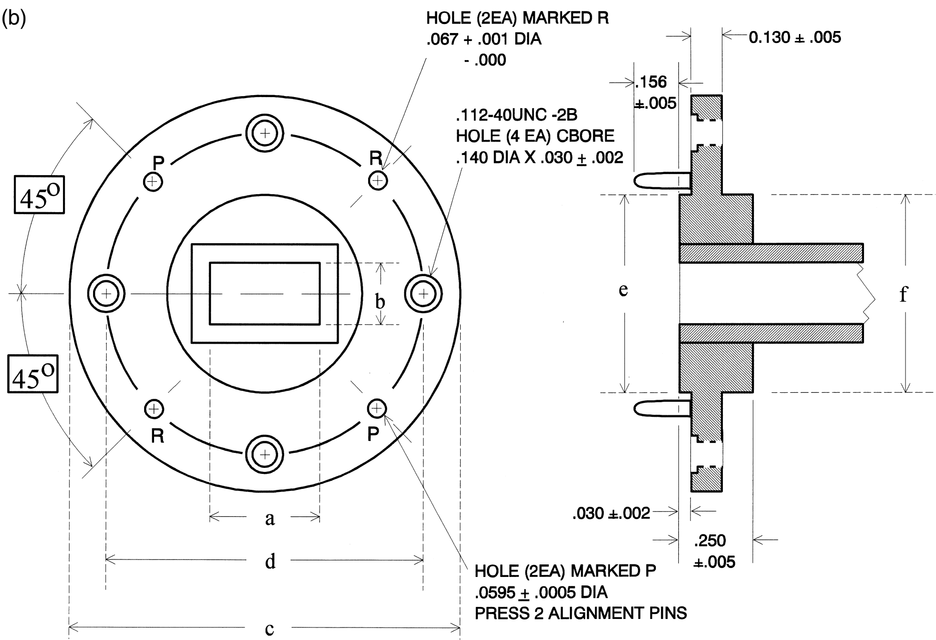


FIGURE B.4 (a) Standard waveguide flange dimensions (rectangular flange); (b) Standard waveguide flange dimensions (circular flange).

TABLE 6 Waveguide Flange Dimensions

Rectangular Flange Dimensions

WR-()	a [inches] (Ref.)	b [inches] (Ref.)	c [inches] ±0.015	d [inches] ±0.015	e [inches] ±0.005	f [inches] ±0.005	G [inches-dia] ±0.0015
28	0.280	0.140	0.750	1.750	0.265	0.250	0.1175
42	0.420	0.170	0.875	0.875	0.335	0.320	0.1175
51	0.510	0.255	1.313	1.313	0.497	0.478	0.1445
62	0.620	0.311	1.313	1.313	0.478	0.497	0.1445
75	0.750	0.375	1.500	2.500	0.560	0.520	0.1445
90	0.900	0.400	1.625	1.625	0.610	0.640	0.1705
112	1.122	0.497	1.875	1.875	0.676	0.737	0.1705

Circular Flange Dimensions

WR-()	a [inches] (Ref.)	b [inches] (Ref.)	c [inches] +0.000 –0.002	d [inches] BSC.	e [inches] ±0.005	f [inches] ±0.005
10	0.1	0.05	0.75	0.5625	0.375	0.312
12	0.122	0.061	0.75	0.5625	0.375	0.312
15	0.148	0.074	0.75	0.5625	0.375	0.312
19	0.188	0.094	1.125	0.9375	0.5	0.468
22	0.224	0.112	1.125	0.9375	0.5	0.468
28	0.28	0.14	1.125	0.9375	0.5	0.468
42	0.42	0.17	1.125	0.9375	0.625	0.625

Coaxial Cables

TABLE 7 Flexible Coax Specifications

RG-()/U	Mil-C-17/()	Zo	Loss dB/100' @1 MHz	Loss dB/100' @10 MHz	Loss dB/100' @100 MHz	Loss dB/100' @1000 MHz	Center Conductor	Outer Conductor	Jacket	Outside Diameter [Inches]	Dielectric	Velocity Factor [%]	Capacitance pF/foot	Dielectric Core Diameter [inches]	Maximum Voltage [RMS]
8		52	0.16	0.56	1.9	7.4	#13 Stranded Bare Copper .058 Dia	Bare Copper Braid, 97%	Black PVC	0.405	Polyethylene	66	29.5	0.285	3700
8A	[3] 163	52	0.16	0.56	1.9	7.4	#13 Stranded Bare Copper .058 Dia	Bare Copper Braid, 97%	Black PVC, Noncontaminating	0.405	Polyethylene	66	29.5	0.285	3700
9		51	0.18	0.62	2.1	8.2	#13 Stranded Silver Coated Copper .086 Dia	Double Braid, Silver Coated Inner, Bare Copper Outer, 97%	Gray PVC, Noncontaminating	0.42	Polyethylene	66	30	0.28	3700
9B	[4] 164	50					#13 Stranded Silver Coated Copper .089 Dia	Double Braid, Silver Coated Inner, Bare Copper Outer, 97%	PVC	0.36	Polyethylene	66		0.285	3701
11	[6]	75	0.19	0.66	2	7.1	#18 Stranded Tinned Copper .048	Bare Copper Braid, 97%	Black PVC	0.405	Flame Retardant Semi-Foam Polyethylene	66	20.5	0.285	300
11A		75	0.19	0.66	2	8.5	#18 Stranded Tinned Copper .048	Bare Copper Braid, 97%	Black PVC, Noncontaminating	0.405	Polyethylene	66	20.5	0.285	3700
58	[28]	50	0.42	1.5	5.4	22.8	#20 Tinned Copper .035 Dia	Tinned Copper Braid, 95%	Black PVC, Noncontaminating	0.193	Polyethylene	66	30.8	0.116	1400
58A		50	0.42	1.5	5.4	22.8	#20 Solid Bare Copper .035 Dia	Tinned Copper Braid, 95%	Black PVC	0.193	Polyethylene	66	30.8	0.116	1400
58C	155	50	0.42	1.5	5.4	22.8	#20 Tinned Copper .035 Dia	Tinned Copper Braid, 95%	Black PVC, Noncontaminating	0.193	Polyethylene	66	30.8	0.116	1400
59	[29]	75	0.6	1.1	3.4	12	#23 Solid Bare Copper Covered Steel .023 Dia	Bare Copper Braid, 95%	Black PVC, Noncontaminating	0.241	Polyethylene	66	20.5	0.146	1700
59B		75	0.6	1.1	3.4	12	#23 Solid Bare Copper Covered Steel .023 Dia	Bare Copper Braid, 95%	Black PVC, Noncontaminating	0.241	Polyethylene	66	20.5	0.146	1700
62A	[30]	93	0.25	0.85	2.7	8.7	#22 Solid Bare Copper Covered Steel .023 Dia	Bare Copper Braid, 95%	Black PVC, Noncontaminating	0.242	Semi-solid Polyethylene	84	13.5	0.146	750

TABLE 7 (continued)

RG-()/U	Mil-C-17/()	Zo	Loss dB/100' @1 MHz	Loss dB/100' @10 MHz	Loss dB/100' @100 MHz	Loss dB/100' @1000 MHz	Center Conductor	Outer Conductor	Jacket	Outside Diameter [Inches]	Dielectric	Velocity Factor [%]	Capacitance pF/foot	Dielectric Core Diameter [inches]	Maximum Voltage [RMS]
62B	[91] 97	93	0.31	0.9	2.9	11	#24 Solid Bare Copper Covered Steel .025 Dia	Bare Copper Braid, 95%	Black PVC, Noncontaminating	0.242	Semi-solid Polyethylene	84	13.5	0.146	750
63	[31]	125	0.19	0.52	1.5	5.8	#22 Solid Bare Copper Covered Steel .025 Dia	Bare Copper Braid, 97%	Black PVC, Noncontaminating	0.405	Semi-solid Polyethylene	84	9.7	0.285	750
71	[90]	93	0.25	0.85	2.7	8.7	#22 Solid Bare Copper Covered Steel .025 Dia	Double Braid, Tinned Copper Outer, Bare Copper Inner, 98%	Black Polyethylene	0.245	Semi-solid Polyethylene	84	13.5	0.146	750
122	[54] 157	50	0.4	1.7	7	29	#22 Stranded Tinned Copper .030 Dia	Tinned Copper Braid, 95%	Black PVC, Noncontaminating	0.16	Polyethylene	66	30.8	0.096	1400
141	[59] 170	50					#18 Solid Silver Coated Copper Covered Steel .037 Dia	Silver Coated Copper Braid, 94%	Fluorinated Ethylene-Propylene	0.17	TFE Teflon	69.5		0.116	1400
141A		50	0.34	1.1	3.9	13.5	#18 Solid Silver Coated Copper Covered Steel .037 Dia	Silver Coated Copper Braid, 94%	Tinted Brown Fiberglass	0.187	TFE Teflon	69.5	29.2	0.116	1400
142	[60] 158	50	0.34	1.1	3.9	13.5	#18 Solid Silver Coated Copper Covered Steel .037 Dia	Double Silver Coated Copper Braid, 94%	Tinted Brown Fluorinated Ethylene-Propylene	0.187	TFE Teflon	69.5	29.2	0.116	1400
174	[119] 173	50	1.9	3.3	8.4	34	#26 Stranded Bare Copper Covered Steel .019 Dia	Tinned Copper Braid, 90%	Black PVC Jacket	0.11	Polyethylene	66	30.8	0.059	1100
178	[93] 169	50					#30 Stranded Silver Coated Copper Covered Steel .012 Dia	Silver Coated Copper Braid, 96%	Fluorinated Ethylene-Propylene	0.071	TFE Teflon	69.5		0.033	750
178B		50	2.6	5.6	14	46	#30 Solid Silver Coated Copper Covered Steel .012 Dia	Silver Coated Copper Braid, 96%	White Fluorinated Ethylene-Propylene	0.071	TFE Teflon	69.5	29.2	0.033	750

179	[94]	75	3	5.3	10	24	#30 Solid Silver Coated Copper Covered Steel .012 Dia	Silver Coated Copper Braid, 95%	Tinted Brown Fluorinated Ethylene-Propylene	0.1	TFE Teflon	69.5	19.5	0.062	900
180	[95]	95	2.4	3.3	5.7	17	#30 Solid Silver Coated Copper Covered Steel .012 Dia	Silver Coated Copper Braid, 95%	Tinted Brown Fluorinated Ethylene-Propylene	0.141	TFE Teflon	69.5	15.4	0.102	1100
187	[68] 94	75					#30 Solid Silver Coated Copper Covered Steel .012 Dia	Silver Coated Copper Braid, 92.3%	Fluorinated Ethylene-Propylene	0.1	TFE Teflon	69.5		0.063	900
212	[73] 162	50	0.26	0.83	2.7	9.8	#15.5 Solid Silver Coated Copper .0556 Dia	Double Silver Coated Copper Braid, 95%	Black PVC Noncontaminating	0.332	Polyethylene	66	30.8	0.185	2200
213	[74] 163	50	0.18	0.62	2.1	8.2	#13 Stranded Bare Copper .089 Dia	Bare Copper Braid 97%	Black PVC Noncontaminating	0.405	Polyethylene	66	30.8	0.285	3700
214	[75] 164	50	0.17	0.55	1.9	8	#13 Stranded Silver Coated Copper .089 Dia	Double Silver Coated Copper Braid, 97%	Black PVC Noncontaminating	0.425	Polyethylene	66	30.8	0.285	3700
216	[77]	75	0.19	0.66	2	7.1	#18 Stranded Tinned Copper .048	Double Bare Copper Braid 95%	Black PVC Noncontaminating	0.425	Polyethylene	66	20.5	0.285	3700
223	[84]	50	0.35	1.2	4.1	14.5	#19 Solid Silver Coated Copper .034 Dia	Double Silver Coated Copper Braid, 95%	Black PVC Noncontaminating	0.212	Polyethylene	66	30.8	0.117	1700
303	[111] 170	50	0.34	1.1	3.9	13.5	#18 Solid Silver Coated Copper Covered Steel .037 Dia	Silver Coated Copper Braid, 95%	Tinted Brown Fluorinated Ethylene-Propylene	0.17	TFE Teflon	69.5	29.2	0.116	1400
316	[113] 172	50	1.2	2.7	8.3	29	#26 Stranded Silver Coated Copper Covered Steel .020 Dia	Silver Coated Copper Braid, 95%	White Fluorinated Ethylene-Propylene	0.098	TFE Teflon	69.5	29.2	0.06	900

Note: Mil-C-17/() part numbers were revised. Initial specification numbers are shown in brackets, current specification numbers are unbracketed.

Source:

Mil-C-17G

Mil-C-17G Supplement 1

Belden Master Catalog, Belden Wire & Cable Co, Richmond, IN.

TABLE 8 Semi-Rigid Coax Specifications

RG-()/U	Mil-C-17 Part Number M17/	Zo	Loss [dB/100 ft] Power [W] @500 MHz	Loss [dB/100 ft] Power [W] @1 GHz	Loss [dB/100 ft] Power [W] @3 GHz	Loss [dB/100 ft] Power [W] @5 GHz	Loss [dB/100 ft] Power [W] @10 GHz	Loss [dB/100 ft] Power [W] @18 GHz	Loss [dB/100 ft] Power [W] @20 GHz	Center Conductor Material	Center Conductor Diameter	Outer Conductor	Outside Diameter [Inches]	Dielectric	Dielectric Constant (1 GHz)	Max Capacitance pF/foot	Dielectric Diameter [inches]	Maximum Voltage (60Hz) [RMS]
	154-00001	50 ± 3.0	42 14	60 10	100 6	140 4.5	190 3.1		280 2	Silver Plated Copper Coated Steel	0.008 ±0.0005	Copper	0.034 ±0.001	Solid PTFE	2.03	29.9	0.026 ±0.001	750
	154-00002	50 ± 3.0	42 14	60 10	100 6	140 4.5	190 3.1		280 2	Silver Plated Copper Coated Steel	0.008 ±0.0005	Tin-Plated Copper	0.034 ±0.002	Solid PTFE	2.03	29.9	0.026 ±0.001	750
	151-00001	50 ± 2.5	28 45	40 32	70 18	90 13	130 9		190 65	Silver Plated Copper Coated Steel	0.0113 ±0.0005	Copper	0.047 ±0.001	Solid PTFE	2.03	29.9	0.037 ±0.001	1000
	151-00002	50 ± 2.5	28 45	40 32	70 18	90 13	130 9		190 6.5	Silver Plated Copper Coated Steel	0.0113 ±0.0005	Tin-Plated Copper	0.047 +0.002 −0.001	Solid PTFE	2.03	29.9	0.037 ±0.001	1000
405	133-RG-405	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Copper Coated Steel	0.0201 ±0.0005	Copper	0.0865 ±0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00001	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Copper Coated Steel	0.0201 ±0.0005	Tin-Plated Copper	0.0865 +0.002 −0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00002	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Copper	0.0201 ±0.0005	Copper	0.0865 ±0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00003	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Copper	0.0201 ±0.0005	Tin-Plated Copper	0.0865 +0.002 −0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00004	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Nickel Copper Coated Steel	0.0201 ±0.0005	Copper	0.0865 ±0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00005	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Nickel Copper Coated Steel	0.0201 ±0.0005	Tin-Plated Copper	0.0865 +0.002 −0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-0006	50 ± 1.5	15 180	22 130		50 54	80 35		130 20	Silver Plated Copper Coated Steel	0.0201 ±0.0005	Copper	0.0865 ±0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000

	133-00007	50 ± 1.5	15 180	22 130		50 54	80 35	130 20	Silver Plated Copper Coated Steel	0.0201 ±0.0005	Tin-Plated Copper	0.086 +0.0021 −0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00008	50 ± 1.5	15 180	22 130		50 54	80 35	130 20	Silver Plated Copper	0.0201 ±0.0005	Copper	0.0865 ±0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00009	50 ± 1.5	15 180	22 130		50 54	80 35	130 20	Silver Plated Copper	0.0201 ±0.0005	Tin-Plated Copper	0.0865 +0.002 −0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00010	50 ± 1.5	15 180	22 130		50 54	80 35	130 20	Silver Plated Nickel Copper Coated Steel	0.0201 ±0.0005	Copper	0.086 ±0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
	133-00011	50 ± 1.5	15 180	22 130		50 54	80 35	130 20	Silver Plated Nickel Copper Coated Steel	0.0201 ±0.0005	Tin-Plated Copper	0.086 +0.002 −0.001	Solid PTFE	2.03	29.9	0.066 ±0.002	5000
402	130-RG-402	50 ± 1.0	8 600	12 450	21 250	29 180	45 120	70 70	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Copper	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00001	50 ± 1.0	8 600	12 450	21 250	29 180	45 120	70 70	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Tin-Plated Copper	0.141 +0.002 −0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00002	50 ± 1.0	8 601	12 451	21 251	29 181	45 121	70 71	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Copper	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00003	50 ± 1.0	8 602	12 452	21 252	29 182	45 122	70 72	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Tin-Plated Copper	0.141 +0.002 −0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00004	50 ± 1.0	8 603	12 453	21 253	29 183	45 123	70 73	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Copper	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00005	50 ± 1.0	8 604	12 454	21 254	29 184	45 124	70 74	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Tin-Plated Copper	0.141 +0.002 −0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00006	50 ± 1.0	8 605	12 455	21 255	29 185	45 125	70 75	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Copper	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000

TABLE 8 (continued)

RG-()/U	Mil-C-17 Part Number M17/	Zo	Loss [dB/100 ft] Power [W] @500 MHz	Loss [dB/100 ft] Power [W] @1 GHz	Loss [dB/100 ft] Power [W] @3 GHz	Loss [dB/100 ft] Power [W] @5 GHz	Loss [dB/100 ft] Power [W] @10 GHz	Loss [dB/100 ft] Power [W] @18 GHz	Loss [dB/100 ft] Power [W] @20 GHz	Center Conductor Material	Center Conductor Diameter	Outer Conductor	Outside Diameter [Inches]	Dielectric	Dielectric Constant (1 GHz)	Max Capacitance pF/foot	Dielectric Diameter [inches]	Maximum Voltage (60Hz) [RMS]
	130-00007	50 ± 1.0	8 606	12 456	21 256	29 186	45 126		70 76	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Tin-Plated Copper	0.141 +0.002 −0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00008	50 ± 1.0	8 607	12 457	21 257	29 187	45 127		70 77	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Aluminum	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00009	50 ± 1.0	8 608	12 458	21 258	29 188	45 128		70 78	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Aluminum	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00010	50 ± 1.0	8 609	12 459	21 259	29 189	45 129		70 79	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Aluminum	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00011	50 ± 1.0	8 610	12 460	21 260	29 190	45 130		70 80	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Aluminum	0.141 ±0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00012	50 ± 1.0	8 611	12 461	21 261	29 191	45 131		70 81	Silver Plated Copper Coated Steel	0.0362 ±0.0007	Silver Plated Copper	0.141 +0.002 −0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
	130-00013	50 ± 1.0	8 612	12 462	21 262	29 192	45 132		70 82	Silver Plated Nickel Copper Coated Steel	0.0362 ±0.0007	Silver Plated Copper	0.141 +0.002 −0.001	Solid PTFE	2.03	29.9	0.1175 ±0.001	5000
401	129-RG-401	50 ± 0.5	4.5 1900	7.5 1400	11 750		33 350	48 200	— —	Silver Plated Copper	0.0641 ±0.001	Copper	0.250 ±0.001	Solid PTFE	2.03	29.9	0.209 ±0.002	7500
	129-00001	50 ± 0.5	4.5 1900	7.5 1400	11 750		33 350	48 200	— —	Silver Plated Copper	0.0641 ±0.001	Tin-Plated Copper	0.250 +0.002 −0.001	Solid PTFE	2.03	29.9	0.209 ±0.002	7500

Notes: Attenuation/Power Ratings are maximum values for families.
Sources: Mil-C-17/130E
Semi-Rigid Coaxial Cable Catalog, Micro-Coax Components, Inc, Collegeville, PA.

Guide To Use of Dissimilar Metals In Sea Water, Marine Atmosphere, and Industrial Atmosphere

Code	Metals	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T
A	Magnesium	G	I / C C	I / C C	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I
B	Zinc Zinc Coating		G	C / C C	I / C C	I / C C	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I
C	Cadmium Beryllium			G	I / C C	I / C C	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I
D	Aluminum, Aluminum-Mg, Aluminum-Zn				G	C / C C	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I
E	Aluminum-Copper					G	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I
F	Steels - Carbon, Low Alloy						G	I / I C	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / C C	I / I I	I / I I	I / I I	I / I I
G	Lead							G	C / C C	I / C C	I / C C	I / C C	I / C C	I / I I	I / I C	I / I C	I / I C	I / I C	I / C C	I / I C	I / I C
H	Tin, Tin-Lead Indium								G	I / I I	I / I C	I / I C	I / C C	I / C C	I / C C	I / C C	C / C C	I / C C	C / C C	I / I I	I / I I
I	Stainless Steels - Martensitic, Ferritic									G	I / I C	I / I I	I / I I	I / I C	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I	I / I I
J	Chromium, Molybdenum, Tungsten										G	C / C C	I / C C	I / I C	I / I I	I / C C	I / C C	I / C C	I / C C	I / C C	I / C C
K	Stainless Steels - Austenitic, Type PH Super Strength Heat Resistant											G	I / I I	I / I I	I / I I	I / C C	I / C C	I / C C	I / C C	I / C C	I / C C
L	Brass-Lead, Bronze												G	I / C C	I / C C	I / I C	I / C C	I / I C	I / C C	I / C C	I / C C
M	Brass - Low Copper, Bronze - Low Copper													G	I / C C	I / C C	I / C C	I / I C	I / C C	I / I C	I / I C
N	Brass - High Copper, Bronze - High Copper														G	I / I C	I / C C	I / I C	I / C C	I / I C	I / I C
O	Copper - High Nickel, Monel															G	I / C C	I / C C	I / C C	I / C C	I / C C
P	Nickel, Cobalt																G	I / C C	I / C C	I / C C	I / C C
Q	Titanium																	G	I / C C	I / C C	I / C C
R	Silver																		G	C / C C	I / C C
S	Palladium, Rhodium, Gold, Platinum																			G	C / C C
T	Graphite																				G

Sea Water
Marine Atmosphere Industrial Atmosphere

Notes:
1. Metals are ranked from most anodic (A) to most cathodic (T).
2. "G" indicates bare identical metal couple compatible in sea water, marine atmosphere, and industrial atmosphere.
3. I indicates joined metals incompatible without appropriate protective coatings in the indicated environment.
4. C indicates joined metals are compatible in the indicated environment.
5. Reference: MIL-STD-889B.

Example: Can High-Copper Brass be put safely in direct contact with Gold?
High-Copper Brass has code letter M, and Gold has code letter T.
Find the intersection of Row M and Column T, to find that these metals are incompatible in salt water and marine atmosphere, but compatible in an industrial atmosphere.

Single Sideband and Image Reject Mixers

TABLE 10 Maximum Tolerable Phase Error (Degrees) for Single Sideband and Image Reject Mixers as a Function of Suppression and Amplitude Imbalance

Amplitude Imbalance [dB]	Suppression												
	−10 dBc	−13 dBc	−15 dBc	−17 dBc	−20 dBc	−23 dBc	−25 dBc	−27 dBc	−30 dBc	−33 dBc	−35 dBc	−37 dBc	−40 dBc
0.00	35.10	25.24	20.17	16.08	11.42	8.10	6.44	5.12	3.62	2.56	2.04	1.62	1.15
0.05	35.10	25.24	20.16	16.08	11.42	8.09	6.43	5.10	3.61	2.54	2.01	1.58	1.10
0.10	35.09	25.23	20.16	16.07	11.40	8.07	6.40	5.07	3.56	2.48	1.93	1.48	0.94
0.15	35.08	25.22	20.14	16.05	11.38	8.04	6.36	5.02	3.48	2.37	1.78	1.28	0.58
0.20	35.08	25.21	20.13	16.03	11.35	7.99	6.30	4.94	3.37	2.20	1.55	0.94	
0.25	35.06	25.19	20.10	16.00	11.30	7.93	6.22	4.84	3.23	1.97	1.20		
0.30	35.05	25.17	20.07	15.96	11.25	7.86	6.13	4.72	3.04	1.63	0.49		
0.35	35.03	25.14	20.04	15.92	11.19	7.77	6.01	4.57	2.79	1.12			
0.40	35.01	25.11	20.00	15.87	11.12	7.66	5.87	4.38	2.48				
0.45	34.99	25.07	19.96	15.81	11.03	7.54	5.72	4.17	2.08				
0.50	34.96	25.04	19.91	15.75	10.94	7.40	5.53	3.91	1.50				
0.55	34.93	24.99	19.85	15.68	10.84	7.25	5.32	3.61					
0.60	34.90	24.95	19.79	15.60	10.72	7.07	5.08	3.25					
0.65	34.87	24.90	19.73	15.51	10.60	6.88	4.81	2.80					
0.70	34.83	24.84	19.65	15.42	10.46	6.66	4.50	2.21					
0.75	34.79	24.78	19.58	15.32	10.31	6.42	4.13	1.32					
0.80	34.75	24.72	19.49	15.21	10.15	6.16	3.70						
0.85	34.70	24.65	19.41	15.10	9.97	5.86	3.18						
0.90	34.66	24.58	19.31	14.98	9.78	5.53	2.52						
0.95	34.61	24.50	19.21	14.84	9.58	5.16	1.53						
1.00	34.55	24.42	19.10	14.70	9.35	4.73							
1.10	34.44	24.24	18.87	14.40	8.86	3.65							
1.20	34.31	24.05	18.62	14.06	8.28	1.83							
1.30	34.17	23.84	18.34	13.67	7.61								
1.40	34.02	23.61	18.03	13.25	6.80								
1.50	33.86	23.35	17.69	12.78	5.82								
1.60	33.68	23.08	17.32	12.25	4.53								
1.70	33.50	22.79	16.92	11.67	2.52								
1.80	33.30	22.48	16.48	11.01									
1.90	33.09	22.14	16.00	10.28									
2.00	32.86	21.78	15.49	9.44									
2.10	32.63	21.39	14.93	8.47									
2.20	32.37	20.98	14.31	7.31									
2.30	32.11	20.54	13.64	5.87									
2.40	31.83	20.07	12.90	3.81									
2.50	31.54	19.56	12.08										
2.60	31.23	19.03	11.17										
2.70	30.90	18.45	10.13										
2.80	30.56	17.83	8.93										
2.90	30.21	17.17	7.48										
3.00	29.83	16.46	5.59										
3.10	29.44	15.68	2.41										
3.20	29.03	14.84											
3.30	28.60	13.92											
3.40	28.16	12.91											
3.50	27.69	11.77											
3.60	27.19	10.47											
3.70	26.68	8.94											
3.80	26.14	7.02											
3.90	25.57	4.24											
4.00	24.98												
4.10	24.35												
4.20	23.69												
4.30	23.00												
4.40	22.27												
4.50	21.49												
4.60	20.67												
4.70	19.79												
4.80	18.86												
4.90	17.85												
5.00	16.76												

Notes: Example: An image reject mixer requires 25 dB of unwanted image suppression, and a maximum amplitude imbalance of 1 dB is expected between the channels. Looking at the intersection of the 1 dB row and –25 dBc column shows that a maximum interchannel phase imbalance of 9.35 degrees is tolerable.
Suppression [dBc] = 10 log ((1 – 2 a cos(p) + a^2) / (1 + 2 a cos(p) + a^2)) where a is the voltage ratio amplitude imbalance and p is the phase imbalance.

Power, Voltage and Decibels

TABLE 11 Power, Voltage, dB Conversion Table

dB	Power Ratio	Voltage Ratio	dB	Power Ratio	Voltage Ratio	dB	Power Ratio	Voltage Ratio	dB	Power Ratio	Voltage Ratio
0	1	1	9	7.943282347	2.818382931	34	2511.886432	50.11872336	60	1000000	1000
0.1	1.023292992	1.011579454	9.5	8.912509381	2.985382619	35	3162.27766	56.23413252	61	1258925.412	1122.018454
0.2	1.047128548	1.023292992	10	10	3.16227766	36	3981.071706	63.09573445	62	1584893.192	1258.925412
0.3	1.071519305	1.035142167	11	12.58925412	3.548133892	37	5011.872336	70.79457844	63	1995262.315	1412.537545
0.4	1.096478196	1.047128548	12	15.84893192	3.981071706	38	6309.573445	79.43282347	64	2511886.432	1584.893192
0.5	1.122018454	1.059253725	13	19.95262315	4.466835922	39	7943.282347	89.12509381	65	3162277.66	1778.27941
0.6	1.148153621	1.071519305	14	25.11886432	5.011872336	40	10000	100	66	3981071.706	1995.262315
0.7	1.174897555	1.083926914	15	31.6227766	5.623413252	41	12589.25412	112.2018454	67	5011872.336	2238.721139
0.8	1.202264435	1.096478196	16	39.81071706	6.309573445	42	15848.93192	125.8925412	68	6309573.445	2511.886432
0.9	1.230268771	1.109174815	17	50.11872336	7.079457844	43	19952.62315	141.2537545	69	7943282.347	2818.382931
1	1.258925412	1.122018454	18	63.09573445	7.943282347	44	25118.86432	158.4893192	70	10000000	3162.27766
1.5	1.412537545	1.188502227	19	79.43282347	8.912509381	45	31622.7766	177.827941	71	12589254.12	3548.133892
2	1.584893192	1.258925412	20	100	10	46	39810.71706	199.5262315	72	15848931.92	3981.071706
2.5	1.77827941	1.333521432	21	125.8925412	11.22018454	47	50118.72336	223.8721139	73	19952623.15	4466.835922
3	1.995262315	1.412537545	22	158.4893192	12.58925412	48	63095.73445	251.1886432	74	25118864.32	5011.872336
3.5	2.238721139	1.496235656	23	199.5262315	14.12537545	49	79432.82347	281.8382931	75	31622776.6	5623.413252
4	2.511886432	1.584893192	24	251.1886432	15.84893192	50	100000	316.227766	76	39810717.06	6309.573445
4.5	2.818382931	1.678804018	25	316.227766	17.7827941	51	125892.5412	354.8133892	77	50118723.36	7079.457844
5	3.16227766	1.77827941	26	398.1071706	19.95262315	52	158489.3192	398.1071706	78	63095734.45	7943.282347
5.5	3.548133892	1.883649089	27	501.1872336	22.38721139	53	199526.2315	446.6835922	79	79432823.47	8912.509381
6	3.981071706	1.995262315	28	630.9573445	25.11886432	54	251188.6432	501.1872336	80	100000000	10000
6.5	4.466835922	2.11348904	29	794.3282347	28.18382931	55	316227.766	562.3413252	81	125892541.2	11220.18454
7	5.011872336	2.238721139	30	1000	31.6227766	56	398107.1706	630.9573445	82	158489319.2	12589.25412
7.5	5.623413252	2.371373706	31	1258.925412	35.48133892	57	501187.2336	707.9457844	83	199526231.5	14125.37545
8	6.309573445	2.511886432	32	1584.893192	39.81071706	58	630957.3445	794.3282347	84	251188643.2	15848.93192
8.5	7.079457844	2.66072506	33	1995.262315	44.66835922	59	794328.2347	891.2509381	85	316227766	17782.7941
86	398107170.6	19952.62315	−5.5	0.281838293	0.530884444	−35	0.000316228	0.017782794	−69	1.25893E-07	0.000354813
87	501187233.6	22387.21139	−6	0.251188643	0.501187234	−36	0.000251189	0.015848932	−70	0.0000001	0.000316228
88	630957344.5	25118.86432	−6.5	0.223872114	0.473151259	−37	0.000199526	0.014125375	−71	7.94328E-08	0.000281838
89	794328234.7	28183.82931	−7	0.199526231	0.446683592	−38	0.000158489	0.012589254	−72	6.30957E-08	0.000251189
90	1000000000	31622.7766	−7.5	0.177827941	0.421696503	−39	0.000125893	0.011220185	−73	5.01187E-08	0.000223872
91	1258925412	35481.33892	−8	0.158489319	0.398107171	−40	0.0001	0.01	−74	3.98107E-08	0.000199526
92	1584893192	39810.71706	−8.5	0.141253754	0.375837404	−41	7.94328E-05	0.008912509	−75	3.16228E-08	0.000177828
93	1995262315	44668.35922	−9	0.125892541	0.354813389	−42	6.30957E-05	0.007943282	−76	2.51189E-08	0.000158489
94	2511886432	50118.72336	−9.5	0.112201845	0.334965439	−43	5.01187E-05	0.007079458	−77	1.99526E-08	0.000141254
95	3162277660	56234.13252	−10	0.1	0.316227766	−44	3.98107E-05	0.006309573	−78	1.58489E-08	0.000125893
96	3981071706	63095.73445	−11	0.079432823	0.281838293	−45	3.16228E-05	0.005623413	−79	1.25893E-08	0.000112202
97	5011872336	70794.57844	−12	0.063095734	0.251188643	−46	2.51189E-05	0.005011872	−80	0.00000001	0.0001
98	6309573445	79432.82347	−13	0.050118723	0.223872114	−47	1.99526E-05	0.004466836	−81	7.94328E-09	8.91251E-05
99	7943282347	89125.09381	−14	0.039810717	0.199526231	−48	1.58489E-05	0.003981072	−82	6.30957E-09	7.94328E-05
100	10000000000	100000	−15	0.031622777	0.177827941	−49	1.25893E-05	0.003548134	−83	5.01187E-09	7.07946E-05
0	1	1	−16	0.025118864	0.158489319	−50	0.00001	0.003162278	−84	3.98107E-09	6.30957E-05
−0.1	0.977237221	0.988553095	−17	0.019952623	0.141253754	−51	7.94328E-06	0.002818383	−85	3.16228E-09	5.62341E-05
−0.2	0.954992586	0.977237221	−18	0.015848932	0.125892541	−52	6.30957E-06	0.002511886	−86	2.51189E-09	5.01187E-05
−0.3	0.933254301	0.966050879	−19	0.012589254	0.112201845	−53	5.01187E-06	0.002238721	−87	1.99526E-09	4.46684E-05
−0.4	0.912010839	0.954992586	−20	0.01	0.1	−54	3.98107E-06	0.001995262	−88	1.58489E-09	3.98107E-05
−0.5	0.891250938	0.944060876	−21	0.007943282	0.089125094	−55	3.16228E-06	0.001778279	−89	1.25893E-09	3.54813E-05
−0.6	0.87096359	0.933254301	−22	0.006309573	0.079432823	−56	2.51189E-06	0.001584893	−90	0.000000001	3.16228E-05
−0.7	0.851138038	0.922571427	−23	0.005011872	0.070794578	−57	1.99526E-06	0.001412538	−91	7.94328E-10	2.81838E-05
−0.8	0.831763771	0.912010839	−24	0.003981072	0.063095734	−58	1.58489E-06	0.001258925	−92	6.30957E-10	2.51189E-05
−0.9	0.812830516	0.901571138	−25	0.003162278	0.056234133	−59	1.25893E-06	0.001122018	−93	5.01187E-10	2.23872E-05
−1	0.794328235	0.891250938	−26	0.002511886	0.050118723	−60	0.000001	0.001	−94	3.98107E-10	1.99526E-05
−1.5	0.707945784	0.841395142	−27	0.001995262	0.044668359	−61	7.94328E-07	0.000891251	−95	3.16228E-10	1.77828E-05
−2	0.630957344	0.794328235	−28	0.001584893	0.039810717	−62	6.30957E-07	0.000794328	−96	2.51189E-10	1.58489E-05
−2.5	0.562341325	0.749894209	−29	0.001258925	0.035481339	−63	5.01187E-07	0.000707946	−97	1.99526E-10	1.41254E-05
−3	0.501187234	0.707945784	−30	0.001	0.031622777	−64	3.98107E-07	0.000630957	−98	1.58489E-10	1.25893E-05
−3.5	0.446683592	0.668343918	−31	0.000794328	0.028183829	−65	3.16228E-07	0.000562341	−99	1.25893E-10	1.12202E-05
−4	0.398107171	0.630957344	−32	0.000630957	0.025118864	−66	2.51189E-07	0.000501187	−100	1E-10	0.00001
−4.5	0.354813389	0.595662144	−33	0.000501187	0.022387211	−67	1.99526E-07	0.000446684			
−5	0.316227766	0.562341325	−34	0.000398107	0.019952623	−68	1.58489E-07	0.000398107			

Notes:
1. Multiply by appropriate factor.
2. Use voltage factor to convert S-parameters.

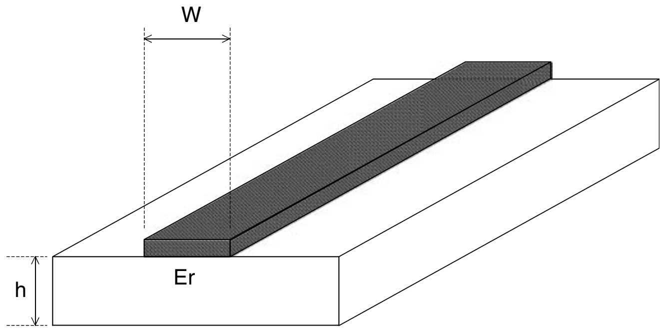


FIGURE B.5

TABLE 12a Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er = 2.2

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	250.3363	2.1861	49.3192	0.0493	57	2.5199	1.8500	0.7959	2.5860
2	123.5739	2.1728	24.5846	0.0983	58	2.4514	1.8471	0.7816	2.6294
3	81.4007	2.1601	16.3417	0.1471	59	2.3854	1.8443	0.7678	2.6727
4	60.3537	2.1480	12.2218	0.1955	60	2.3218	1.8416	0.7544	2.7160
5	47.7490	2.1364	9.7510	0.2438	61	2.2603	1.8389	0.7415	2.7592
6	39.3615	2.1253	8.1046	0.2918	62	2.2010	1.8362	0.7290	2.8024
7	33.3816	2.1146	6.9294	0.3395	63	2.1437	1.8336	0.7170	2.8456
8	28.9051	2.1044	6.0485	0.3871	64	2.0883	1.8310	0.7053	2.8887
9	25.4299	2.0946	5.3639	0.4345	65	2.0347	1.8285	0.6939	2.9318
10	22.6550	2.0851	4.8166	0.4817	66	1.9829	1.8259	0.6829	2.9749
11	20.3889	2.0760	4.3692	0.5287	67	1.9327	1.8235	0.6723	3.0179
12	18.5040	2.0673	3.9967	0.5755	68	1.8919	1.8214	0.6620	3.0612
13	16.9121	2.0589	3.6817	0.6222	69	1.8447	1.8190	0.6520	3.1042
14	15.5503	2.0508	3.4120	0.6688	70	1.7990	1.8166	0.6423	3.1471
15	14.3723	2.0429	3.1784	0.7152	71	1.7548	1.8143	0.6328	3.1900
16	13.3435	2.0354	2.9743	0.7614	72	1.7119	1.8120	0.6236	3.2329
17	12.4375	2.0280	2.7943	0.8075	73	1.6704	1.8097	0.6147	3.2757
18	11.6337	2.0210	2.6344	0.8536	74	1.6301	1.8075	0.6060	3.3186
19	10.9159	2.0141	2.4915	0.8994	75	1.5910	1.8053	0.5976	3.3614
20	10.2711	2.0075	2.3631	0.9452	76	1.5531	1.8031	0.5894	3.4041
21	9.6889	2.0010	2.2469	0.9909	77	1.5163	1.8010	0.5814	3.4469
22	9.1607	1.9948	2.1414	1.0365	78	1.4806	1.7988	0.5736	3.4896
23	8.6793	1.9887	2.0452	1.0819	79	1.4458	1.7968	0.5660	3.5322
24	8.2389	1.9828	1.9571	1.1273	80	1.4121	1.7947	0.5586	3.5749
25	7.8346	1.9771	1.8761	1.1726	81	1.3793	1.7926	0.5514	3.6175
26	7.4621	1.9715	1.8014	1.2177	82	1.3474	1.7906	0.5443	3.6601
27	7.1178	1.9661	1.7323	1.2628	83	1.3164	1.7886	0.5375	3.7027
28	6.7988	1.9608	1.6682	1.3078	84	1.2862	1.7867	0.5308	3.7453
29	6.5024	1.9557	1.6085	1.3528	85	1.2568	1.7847	0.5243	3.7878

TABLE 12a (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
30	6.2263	1.9507	1.5529	1.3976	86	1.2282	1.7828	0.5179	3.8303
31	5.9686	1.9458	1.5010	1.4424	87	1.2003	1.7809	0.5117	3.8728
32	5.7274	1.9410	1.4523	1.4871	88	1.1732	1.7791	0.5056	3.9152
33	5.5013	1.9364	1.4066	1.5318	89	1.1467	1.7772	0.4996	3.9577
34	5.2890	1.9319	1.3636	1.5763	90	1.1210	1.7754	0.4938	4.0001
35	5.0892	1.9274	1.3231	1.6208	91	1.0959	1.7736	0.4882	4.0424
36	4.9009	1.9231	1.2849	1.6653	92	1.0714	1.7718	0.4826	4.0848
37	4.7231	1.9189	1.2488	1.7096	93	1.0476	1.7700	0.4772	4.1272
38	4.5550	1.9147	1.2146	1.7539	94	1.0243	1.7683	0.4719	4.1695
39	4.3959	1.9107	1.1822	1.7982	95	1.0016	1.7665	0.4667	4.2118
40	4.2450	1.9067	1.1515	1.8424	96	0.9795	1.7658	0.4617	4.2552
41	4.1018	1.9028	1.1223	1.8865	97	0.9579	1.7651	0.4569	4.2987
42	3.9658	1.8990	1.0945	1.9306	98	0.9369	1.7644	0.4521	4.3421
43	3.8363	1.8953	1.0680	1.9746	99	0.9163	1.7637	0.4475	4.3855
44	3.7129	1.8917	1.0427	2.0186	100	0.8962	1.7629	0.4429	4.4289
45	3.5953	1.8881	1.0185	2.0625	101	0.8767	1.7621	0.4384	4.4722
46	3.4831	1.8846	0.9955	2.1064	102	0.8575	1.7613	0.4340	4.5154
47	3.3758	1.8811	0.9734	2.1502	103	0.8389	1.7605	0.4297	4.5586
48	3.2733	1.8778	0.9523	2.1940	104	0.8206	1.7596	0.4255	4.6018
49	3.1751	1.8745	0.9320	2.2378	105	0.8028	1.7588	0.4213	4.6449
50	3.0811	1.8712	0.9126	2.2814	106	0.7855	1.7579	0.4172	4.6880
51	2.9909	1.8680	0.8939	2.3251	107	0.7685	1.7570	0.4132	4.7310
52	2.9044	1.8649	0.8760	2.3687	108	0.7519	1.7561	0.4093	4.7740
53	2.8214	1.8618	0.8588	2.4122	109	0.7357	1.7552	0.4054	4.8169
54	2.7416	1.8587	0.8422	2.4557	110	0.7198	1.7543	0.4016	4.8599
55	2.6648	1.8558	0.8262	2.4992	111	0.7043	1.7534	0.3979	4.9027
56	2.5910	1.8528	0.8108	2.5426	112	0.6892	1.7524	0.3943	4.9456
113	0.6744	1.7515	0.3907	4.9884	132	0.4484	1.7330	0.3327	5.7964
114	0.6600	1.7505	0.3871	5.0312	133	0.4389	1.7321	0.3301	5.8387
115	0.6459	1.7496	0.3837	5.0739	134	0.4296	1.7311	0.3275	5.8810
116	0.6320	1.7486	0.3803	5.1166	135	0.4206	1.7302	0.3250	5.9232
117	0.6185	1.7477	0.3769	5.1593	136	0.4117	1.7292	0.3225	5.9654
118	0.6053	1.7467	0.3736	5.2020	137	0.4030	1.7283	0.3201	6.0076
119	0.5924	1.7457	0.3704	5.2446	138	0.3945	1.7273	0.3177	6.0498
120	0.5798	1.7447	0.3672	5.2872	139	0.3862	1.7264	0.3153	6.0920
121	0.5675	1.7438	0.3640	5.3298	140	0.3781	1.7254	0.3130	6.1342
122	0.5554	1.7428	0.3609	5.3723	141	0.3702	1.7245	0.3107	6.1763
123	0.5436	1.7418	0.3579	5.4148	142	0.3624	1.7236	0.3084	6.2185
124	0.5321	1.7408	0.3549	5.4573	143	0.3547	1.7226	0.3062	6.2606
125	0.5208	1.7399	0.3520	5.4998	144	0.3473	1.7217	0.3039	6.3027
126	0.5097	1.7389	0.3491	5.5422	145	0.3400	1.7208	0.3018	6.3448
127	0.4989	1.7379	0.3462	5.5846	146	0.3328	1.7199	0.2996	6.3868
128	0.4884	1.7369	0.3434	5.6270	147	0.3259	1.7190	0.2975	6.4289
129	0.4780	1.7360	0.3407	5.6694	148	0.3190	1.7181	0.2954	6.4709
130	0.4679	1.7350	0.3380	5.7118	149	0.3123	1.7172	0.2934	6.5130
131	0.4580	1.7340	0.3353	5.7541	150	0.3058	1.7163	0.2913	6.5550

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12b Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er =3.78

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	190.5772	3.7382	64.4927	0.0645	26	5.4498	3.1668	2.2831	1.5433
2	93.9049	3.6989	32.0763	0.1283	27	5.1890	3.1537	2.1940	1.5994
3	61.7511	3.6619	21.2770	0.1915	28	4.9474	3.1410	2.1113	1.6553
4	45.7086	3.6271	15.8817	0.2541	29	4.7230	3.1287	2.0345	1.7110
5	36.1035	3.5942	12.6477	0.3162	30	4.5140	3.1167	1.9629	1.7666
6	29.7137	3.5632	10.4941	0.3778	31	4.3190	3.1051	1.8961	1.8221
7	25.1593	3.5337	8.9578	0.4389	32	4.1365	3.0938	1.8335	1.8775
8	21.7508	3.5059	7.8070	0.4997	33	3.9655	3.0827	1.7747	1.9327
9	19.1053	3.4794	6.9133	0.5600	34	3.8050	3.0720	1.7195	1.9878
10	16.9935	3.4542	6.1994	0.6199	35	3.6539	3.0616	1.6676	2.0428
11	15.2694	3.4301	5.6162	0.6796	36	3.5116	3.0514	1.6185	2.0976
12	13.8357	3.4072	5.1309	0.7389	37	3.3773	3.0414	1.5722	2.1524
13	12.6252	3.3853	4.7210	0.7978	38	3.2504	3.0317	1.5284	2.2070
14	11.5899	3.3643	4.3702	0.8566	39	3.1302	3.0222	1.4869	2.2616
15	10.6946	3.3442	4.0666	0.9150	40	3.0164	3.0130	1.4475	2.3160
16	9.9129	3.3249	3.8014	0.9732	41	2.9083	3.0039	1.4101	2.3703
17	9.2247	3.3064	3.5678	1.0311	42	2.8056	2.9951	1.3745	2.4246
18	8.6143	3.2885	3.3605	1.0888	43	2.7080	2.9864	1.3406	2.4787
19	8.0694	3.2714	3.1754	1.1463	44	2.6150	2.9780	1.3082	2.5327
20	7.5800	3.2549	3.0090	1.2036	45	2.5263	2.9697	1.2774	2.5867
21	7.1382	3.2389	2.8586	1.2607	46	2.4417	2.9615	1.2479	2.6406
22	6.7375	3.2235	2.7222	1.3175	47	2.3609	2.9536	1.2197	2.6943
23	6.3724	3.2086	2.5978	1.3743	48	2.2836	2.9458	1.1927	2.7480
24	6.0386	3.1942	2.4840	1.4308	49	2.2097	2.9381	1.1669	2.8016
25	5.7321	3.1803	2.3794	1.4871	50	2.1389	2.9306	1.1421	2.8552
51	2.0711	2.9233	1.1183	2.9086	101	0.5140	2.7126	0.5439	5.5488
52	2.0060	2.9160	1.0954	2.9620	102	0.5008	2.7099	0.5383	5.6008
53	1.9435	2.9089	1.0734	3.0152	103	0.4878	2.7071	0.5328	5.6529
54	1.8840	2.9020	1.0523	3.0685	104	0.4752	2.7043	0.5274	5.7048
55	1.8268	2.8952	1.0320	3.1217	105	0.4630	2.7016	0.5222	5.7568
56	1.7719	2.8886	1.0124	3.1748	106	0.4511	2.6989	0.5170	5.8086
57	1.7190	2.8820	0.9935	3.2278	107	0.4394	2.6961	0.5119	5.8605
58	1.6682	2.8756	0.9752	3.2807	108	0.4281	2.6934	0.5069	5.9123
59	1.6192	2.8693	0.9577	3.3336	109	0.4171	2.6907	0.5020	5.9640
60	1.5720	2.8631	0.9407	3.3865	110	0.4064	2.6880	0.4972	6.0157
61	1.5265	2.8569	0.9243	3.4392	111	0.3960	2.6853	0.4924	6.0674
62	1.4826	2.8509	0.9084	3.4919	112	0.3858	2.6826	0.4878	6.1190
63	1.4402	2.8450	0.8931	3.5446	113	0.3759	2.6800	0.4832	6.1706
64	1.3993	2.8392	0.8782	3.5971	114	0.3663	2.6774	0.4788	6.2221
65	1.3597	2.8334	0.8638	3.6496	115	0.3569	2.6747	0.4744	6.2736
66	1.3215	2.8278	0.8499	3.7021	116	0.3477	2.6721	0.4701	6.3251
67	1.2845	2.8222	0.8364	3.7545	117	0.3388	2.6696	0.4658	6.3765
68	1.2488	2.8167	0.8233	3.8068	118	0.3301	2.6670	0.4616	6.4279
69	1.2142	2.8113	0.8106	3.8591	119	0.3217	2.6644	0.4575	6.4793
70	1.1807	2.8060	0.7982	3.9113	120	0.3134	2.6619	0.4535	6.5307
71	1.1482	2.8008	0.7862	3.9635	121	0.3054	2.6594	0.4496	6.5820
72	1.1168	2.7956	0.7746	4.0156	122	0.2976	2.6569	0.4457	6.6333
73	1.0863	2.7905	0.7633	4.0676	123	0.2900	2.6544	0.4418	6.6845
74	1.0568	2.7855	0.7523	4.1196	124	0.2826	2.6520	0.4381	6.7358
75	1.0282	2.7805	0.7416	4.1716	125	0.2754	2.6496	0.4344	6.7870
76	1.0004	2.7756	0.7312	4.2235	126	0.2683	2.6472	0.4307	6.8382
77	0.9735	2.7737	0.7215	4.2776	127	0.2615	2.6448	0.4271	6.8893
78	0.9474	2.7717	0.7120	4.3316	128	0.2548	2.6424	0.4236	6.9405
79	0.9220	2.7696	0.7027	4.3855	129	0.2483	2.6401	0.4201	6.9916

TABLE 12b (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
80	0.8974	2.7675	0.6936	4.4393	130	0.2420	2.6378	0.4167	7.0427
81	0.8735	2.7653	0.6848	4.4929	131	0.2358	2.6355	0.4134	7.0938
82	0.8503	2.7630	0.6762	4.5465	132	0.2298	2.6332	0.4101	7.1449
83	0.8277	2.7606	0.6677	4.6000	133	0.2239	2.6309	0.4068	7.1959
84	0.8058	2.7582	0.6595	4.6534	134	0.2182	2.6287	0.4036	7.2469
85	0.7845	2.7557	0.6514	4.7067	135	0.2126	2.6265	0.4004	7.2980
86	0.7639	2.7532	0.6436	4.7599	136	0.2072	2.6243	0.3973	7.3490
87	0.7438	2.7506	0.6359	4.8130	137	0.2019	2.6221	0.3943	7.3999
88	0.7242	2.7481	0.6284	4.8660	138	0.1968	2.6200	0.3912	7.4509
89	0.7052	2.7454	0.6210	4.9190	139	0.1917	2.6179	0.3883	7.5019
90	0.6868	2.7428	0.6138	4.9718	140	0.1869	2.6158	0.3853	7.5528
91	0.6688	2.7401	0.6068	5.0246	141	0.1821	2.6137	0.3825	7.6037
92	0.6513	2.7374	0.5999	5.0773	142	0.1775	2.6116	0.3796	7.6546
93	0.6344	2.7347	0.5931	5.1300	143	0.1729	2.6096	0.3768	7.7055
94	0.6178	2.7320	0.5865	5.1826	144	0.1685	2.6076	0.3741	7.7564
95	0.6018	2.7292	0.5801	5.2351	145	0.1642	2.6056	0.3713	7.8073
96	0.5861	2.7265	0.5737	5.2875	146	0.1600	2.6036	0.3687	7.8582
97	0.5709	2.7237	0.5675	5.3399	147	0.1560	2.6017	0.3660	7.9091
98	0.5561	2.7209	0.5615	5.3922	148	0.1520	2.5998	0.3634	7.9599
99	0.5417	2.7182	0.5555	5.4444	149	0.1481	2.5979	0.3608	8.0108
100	0.5277	2.7154	0.5497	5.4966	150	0.1444	2.5960	0.3583	8.0616

Notes: Calculation of W/H has an error of less than 1%.
Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12c Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er =5.75

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	154.1545	5.6626	79.3758	0.0794	55	1.3098	4.1200	1.2310	3.7239
2	75.8057	5.5818	39.4035	0.1576	56	1.2664	4.1088	1.2074	3.7864
3	49.7547	5.5068	26.0920	0.2348	57	1.2248	4.0978	1.1846	3.8488
4	36.7611	5.4372	19.4449	0.3111	58	1.1847	4.0869	1.1627	3.9112
5	28.9839	5.3723	15.4628	0.3866	59	1.1461	4.0763	1.1415	3.9734
6	23.8117	5.3116	12.8127	0.4613	60	1.1089	4.0658	1.1210	4.0355
7	20.1263	5.2548	10.9235	0.5352	61	1.0731	4.0555	1.1012	4.0976
8	17.3689	5.2014	9.5094	0.6086	62	1.0386	4.0453	1.0821	4.1596
9	15.2296	5.1512	8.4118	0.6814	63	1.0054	4.0353	1.0636	4.2214
10	13.5223	5.1037	7.5357	0.7536	64	0.9733	4.0305	1.0464	4.2859
11	12.1289	5.0589	6.8204	0.8253	65	0.9423	4.0265	1.0297	4.3507
12	10.9706	5.0163	6.2257	0.8965	66	0.9124	4.0222	1.0136	4.4153
13	9.9929	4.9759	5.7236	0.9673	67	0.8836	4.0178	0.9979	4.4797
14	9.1570	4.9375	5.2942	1.0377	68	0.8557	4.0132	0.9827	4.5439
15	8.4343	4.9008	4.9229	1.1077	69	0.8288	4.0084	0.9679	4.6080
16	7.8035	4.8659	4.5987	1.1773	70	0.8028	4.0035	0.9535	4.6719
17	7.2484	4.8324	4.3133	1.2466	71	0.7777	3.9985	0.9394	4.7357
18	6.7562	4.8004	4.0602	1.3155	72	0.7534	3.9933	0.9258	4.7993
19	6.3169	4.7697	3.8342	1.3841	73	0.7299	3.9881	0.9125	4.8628
20	5.9225	4.7403	3.6312	1.4525	74	0.7072	3.9828	0.8996	4.9261
21	5.5666	4.7119	3.4479	1.5205	75	0.6853	3.9774	0.8870	4.9893
22	5.2438	4.6847	3.2817	1.5883	76	0.6641	3.9720	0.8747	5.0524
23	4.9499	4.6584	3.1302	1.6559	77	0.6435	3.9665	0.8628	5.1153
24	4.6812	4.6331	2.9916	1.7232	78	0.6236	3.9609	0.8511	5.1781
25	4.4346	4.6087	2.8644	1.7902	79	0.6044	3.9554	0.8397	5.2408

TABLE 12c (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
26	4.2075	4.5851	2.7471	1.8571	80	0.5858	3.9498	0.8287	5.3034
27	3.9979	4.5623	2.6388	1.9237	81	0.5678	3.9442	0.8178	5.3659
28	3.8037	4.5402	2.5384	1.9901	82	0.5503	3.9386	0.8073	5.4283
29	3.6233	4.5187	2.4451	2.0563	83	0.5334	3.9329	0.7970	5.4905
30	3.4554	4.4980	2.3581	2.1223	84	0.5171	3.9273	0.7870	5.5527
31	3.2988	4.4778	2.2769	2.1881	85	0.5012	3.9217	0.7771	5.6148
32	3.1523	4.4583	2.2010	2.2538	86	0.4859	3.9161	0.7676	5.6768
33	3.0151	4.4393	2.1297	2.3193	87	0.4710	3.9105	0.7582	5.7387
34	2.8863	4.4208	2.0628	2.3846	88	0.4566	3.9049	0.7490	5.8005
35	2.7652	4.4028	1.9997	2.4497	89	0.4427	3.8994	0.7401	5.8623
36	2.6511	4.3853	1.9403	2.5147	90	0.4292	3.8938	0.7314	5.9239
37	2.5434	4.3682	1.8842	2.5795	91	0.4161	3.8883	0.7228	5.9855
38	2.4417	4.3516	1.8311	2.6441	92	0.4034	3.8829	0.7144	6.0470
39	2.3455	4.3353	1.7808	2.7087	93	0.3912	3.8774	0.7063	6.1085
40	2.2543	4.3195	1.7331	2.7730	94	0.3793	3.8720	0.6983	6.1699
41	2.1678	4.3040	1.6878	2.8373	95	0.3677	3.8667	0.6904	6.2312
42	2.0856	4.2889	1.6448	2.9014	96	0.3565	3.8614	0.6828	6.2925
43	2.0075	4.2741	1.6037	2.9653	97	0.3457	3.8561	0.6753	6.3537
44	1.9289	4.2588	1.5645	3.0288	98	0.3352	3.8508	0.6679	6.4148
45	1.8591	4.2449	1.5272	3.0926	99	0.3250	3.8456	0.6607	6.4759
46	1.7926	4.2312	1.4916	3.1562	100	0.3152	3.8405	0.6537	6.5369
47	1.7291	4.2178	1.4576	3.2198	101	0.3056	3.8354	0.6468	6.5979
48	1.6684	4.2048	1.4250	3.2832	102	0.2963	3.8304	0.6400	6.6588
49	1.6104	4.1919	1.3938	3.3464	103	0.2874	3.8254	0.6334	6.7197
50	1.5549	4.1794	1.3638	3.4096	104	0.2787	3.8204	0.6269	6.7806
51	1.5017	4.1671	1.3351	3.4727	105	0.2702	3.8155	0.6205	6.8414
52	1.4507	4.1550	1.3076	3.5356	106	0.2620	3.8107	0.6143	6.9022
53	1.4018	4.1431	1.2811	3.5985	107	0.2541	3.8059	0.6082	6.9629
54	1.3549	4.1315	1.2556	3.6612	108	0.2464	3.8011	0.6022	7.0236
109	0.2389	3.7964	0.5963	7.0842	130	0.1255	3.7101	0.4942	8.3525
110	0.2317	3.7918	0.5905	7.1449	131	0.1217	3.7065	0.4902	8.4127
111	0.2247	3.7872	0.5848	7.2055	132	0.1180	3.7030	0.4863	8.4729
112	0.2179	3.7827	0.5792	7.2660	133	0.1144	3.6996	0.4824	8.5331
113	0.2113	3.7782	0.5738	7.3265	134	0.1110	3.6962	0.4786	8.5933
114	0.2049	3.7738	0.5684	7.3871	135	0.1076	3.6928	0.4748	8.6535
115	0.1987	3.7694	0.5631	7.4475	136	0.1044	3.6895	0.4711	8.7137
116	0.1927	3.7651	0.5580	7.5080	137	0.1012	3.6863	0.4675	8.7739
117	0.1869	3.7608	0.5529	7.5684	138	0.0982	3.6830	0.4639	8.8341
118	0.1813	3.7566	0.5479	7.6288	139	0.0952	3.6799	0.4603	8.8943
119	0.1758	3.7524	0.5430	7.6892	140	0.0923	3.6767	0.4569	8.9544
120	0.1705	3.7483	0.5382	7.7496	141	0.0896	3.6737	0.4534	9.0146
121	0.1653	3.7443	0.5334	7.8099	142	0.0869	3.6706	0.4500	9.0748
122	0.1603	3.7403	0.5288	7.8703	143	0.0842	3.6676	0.4467	9.1350
123	0.1555	3.7363	0.5242	7.9306	144	0.0817	3.6647	0.4434	9.1952
124	0.1508	3.7324	0.5197	7.9909	145	0.0792	3.6618	0.4402	9.2553
125	0.1462	3.7286	0.5153	8.0512	146	0.0768	3.6589	0.4370	9.3155
126	0.1418	3.7248	0.5109	8.1115	147	0.0745	3.6561	0.4339	9.3757
127	0.1375	3.7210	0.5066	8.1717	148	0.0723	3.6533	0.4308	9.4359
128	0.1334	3.7173	0.5024	8.2320	149	0.0701	3.6505	0.4277	9.4961
129	0.1294	3.7137	0.4983	8.2922	150	0.0680	3.6478	0.4247	9.5562

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12d Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er = 9.4

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	120.1221	9.2047	101.2010	0.1012	24	3.4066	7.1750	3.7229	2.1444
2	58.8859	9.0280	50.1125	0.2004	25	3.2161	7.1309	3.5630	2.2269
3	38.5353	8.8676	33.1101	0.2980	26	3.0409	7.0885	3.4157	2.3090
4	28.3901	8.7212	24.6268	0.3940	27	2.8791	7.0475	3.2797	2.3909
5	22.3208	8.5871	19.5493	0.4887	28	2.7293	7.0079	3.1537	2.4725
6	18.2864	8.4635	16.1735	0.5822	29	2.5902	6.9697	3.0366	2.5538
7	15.4132	8.3493	13.7691	0.6747	30	2.4609	6.9326	2.9276	2.6348
8	13.2647	8.2433	11.9712	0.7662	31	2.3403	6.8967	2.8258	2.7156
9	11.5985	8.1445	10.5771	0.8567	32	2.2275	6.8619	2.7306	2.7961
10	10.2695	8.0521	9.4653	0.9465	33	2.1220	6.8281	2.6413	2.8764
11	9.1854	7.9655	8.5584	1.0356	34	2.0229	6.7952	2.5574	2.9564
12	8.2846	7.8841	7.8050	1.1239	35	1.9221	6.7606	2.4780	3.0356
13	7.5248	7.8074	7.1695	1.2116	36	1.8364	6.7301	2.4037	3.1152
14	6.8754	7.7348	6.6264	1.2988	37	1.7556	6.7004	2.3336	3.1947
15	6.3143	7.6661	6.1571	1.3853	38	1.6792	6.6715	2.2673	3.2740
16	5.8248	7.6009	5.7477	1.4714	39	1.6070	6.6434	2.2045	3.3530
17	5.3942	7.5389	5.3875	1.5570	40	1.5386	6.6159	2.1449	3.4319
18	5.0126	7.4798	5.0682	1.6421	41	1.4738	6.5891	2.0884	3.5105
19	4.6722	7.4234	4.7833	1.7268	42	1.4122	6.5629	2.0346	3.5890
20	4.3668	7.3694	4.5276	1.8110	43	1.3537	6.5373	1.9834	3.6673
21	4.0913	7.3178	4.2968	1.8949	44	1.2981	6.5122	1.9346	3.7454
22	3.8416	7.2683	4.0876	1.9784	45	1.2451	6.4877	1.8880	3.8233
23	3.6143	7.2207	3.8971	2.0616	46	1.1946	6.4638	1.8436	3.9010
47	1.1465	6.4403	1.8011	3.9786	99	0.1530	5.8352	0.8139	7.9771
48	1.1005	6.4173	1.7604	4.0560	100	0.1472	5.8268	0.8052	8.0518
49	1.0567	6.3948	1.7215	4.1332	101	0.1417	5.8184	0.7966	8.1265
50	1.0148	6.3728	1.6841	4.2103	102	0.1365	5.8102	0.7883	8.2012
51	0.9747	6.3596	1.6494	4.2901	103	0.1314	5.8021	0.7801	8.2758
52	0.9364	6.3507	1.6165	4.3711	104	0.1265	5.7942	0.7720	8.3504
53	0.8997	6.3412	1.5849	4.4519	105	0.1217	5.7864	0.7642	8.4251
54	0.8646	6.3312	1.5543	4.5323	106	0.1172	5.7787	0.7565	8.4997
55	0.8309	6.3208	1.5248	4.6124	107	0.1128	5.7712	0.7489	8.5742
56	0.7987	6.3100	1.4963	4.6923	108	0.1086	5.7638	0.7415	8.6488
57	0.7678	6.2989	1.4687	4.7719	109	0.1046	5.7565	0.7342	8.7234
58	0.7382	6.2875	1.4421	4.8512	110	0.1006	5.7493	0.7271	8.7979
59	0.7098	6.2759	1.4163	4.9303	111	0.0969	5.7423	0.7201	8.8725
60	0.6826	6.2641	1.3914	5.0091	112	0.0933	5.7354	0.7133	8.9470
61	0.6564	6.2521	1.3673	5.0877	113	0.0898	5.7286	0.7065	9.0216
62	0.6313	6.2400	1.3439	5.1661	114	0.0864	5.7219	0.6999	9.0961
63	0.6072	6.2278	1.3213	5.2443	115	0.0832	5.7154	0.6934	9.1707
64	0.5841	6.2156	1.2994	5.3223	116	0.0801	5.7090	0.6871	9.2452
65	0.5619	6.2032	1.2781	5.4001	117	0.0771	5.7026	0.6808	9.3197
66	0.5406	6.1909	1.2575	5.4777	118	0.0743	5.6964	0.6747	9.3943
67	0.5201	6.1786	1.2375	5.5552	119	0.0715	5.6903	0.6687	9.4688
68	0.5004	6.1662	1.2181	5.6325	120	0.0688	5.6843	0.6627	9.5434
69	0.4814	6.1539	1.1992	5.7096	121	0.0662	5.6785	0.6569	9.6179
70	0.4632	6.1417	1.1809	5.7866	122	0.0638	5.6727	0.6512	9.6925
71	0.4457	6.1295	1.1631	5.8634	123	0.0614	5.6670	0.6456	9.7670
72	0.4289	6.1173	1.1458	5.9401	124	0.0591	5.6615	0.6401	9.8416
73	0.4127	6.1053	1.1290	6.0166	125	0.0569	5.6560	0.6346	9.9162
74	0.3972	6.0933	1.1127	6.0931	126	0.0548	5.6506	0.6293	9.9907
75	0.3822	6.0814	1.0968	6.1694	127	0.0527	5.6454	0.6241	10.0653
76	0.3679	6.0696	1.0813	6.2456	128	0.0508	5.6402	0.6189	10.1399
77	0.3540	6.0579	1.0662	6.3217	129	0.0489	5.6351	0.6138	10.2146
78	0.3407	6.0464	1.0516	6.3977	130	0.0471	5.6301	0.6088	10.2892
79	0.3279	6.0350	1.0373	6.4736	131	0.0453	5.6252	0.6039	10.3638

TABLE 12d (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
80	0.3156	6.0236	1.0233	6.5494	132	0.0436	5.6204	0.5991	10.4385
81	0.3038	6.0125	1.0098	6.6251	133	0.0420	5.6157	0.5943	10.5131
82	0.2924	6.0014	0.9965	6.7007	134	0.0404	5.6111	0.5897	10.5878
83	0.2815	5.9905	0.9836	6.7762	135	0.0389	5.6065	0.5850	10.6625
84	0.2709	5.9797	0.9710	6.8517	136	0.0375	5.6021	0.5805	10.7372
85	0.2608	5.9691	0.9588	6.9271	137	0.0361	5.5977	0.5761	10.8119
86	0.2510	5.9586	0.9468	7.0024	138	0.0347	5.5934	0.5717	10.8867
87	0.2416	5.9482	0.9351	7.0777	139	0.0334	5.5892	0.5673	10.9614
88	0.2326	5.9380	0.9237	7.1529	140	0.0322	5.5850	0.5631	11.0362
89	0.2239	5.9280	0.9125	7.2281	141	0.0310	5.5809	0.5589	11.1110
90	0.2155	5.9180	0.9016	7.3032	142	0.0298	5.5770	0.5547	11.1858
91	0.2074	5.9083	0.8910	7.3782	143	0.0287	5.5730	0.5507	11.2606
92	0.1997	5.8986	0.8806	7.4532	144	0.0276	5.5692	0.5467	11.3354
93	0.1922	5.8891	0.8704	7.5281	145	0.0266	5.5654	0.5427	11.4103
94	0.1850	5.8798	0.8605	7.6031	146	0.0256	5.5617	0.5388	11.4852
95	0.1781	5.8706	0.8507	7.6779	147	0.0247	5.5581	0.5350	11.5600
96	0.1715	5.8616	0.8412	7.7528	148	0.0237	5.5545	0.5312	11.6350
97	0.1651	5.8526	0.8319	7.8276	149	0.0229	5.5510	0.5274	11.7099
98	0.1589	5.8439	0.8228	7.9023	150	0.0220	5.5476	0.5238	11.7848

Notes: Calculation of W/H has an error of less than 1%.

TABLE 12e Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er = 9.8

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	117.6023	9.5914	103.3045	0.1033	57	0.7347	6.5379	1.4963	4.8615
2	57.6329	9.4030	51.1424	0.2046	58	0.7059	6.5254	1.4691	4.9421
3	37.7044	9.2322	33.7840	0.3041	59	0.6783	6.5128	1.4428	5.0224
4	27.7700	9.0767	25.1237	0.4020	60	0.6519	6.5000	1.4174	5.1025
5	21.8272	8.9344	19.9408	0.4985	61	0.6265	6.4870	1.3927	5.1824
6	17.8771	8.8035	16.4952	0.5938	62	0.6022	6.4740	1.3689	5.2621
7	15.0641	8.6827	14.0413	0.6880	63	0.5788	6.4609	1.3458	5.3415
8	12.9606	8.5706	12.2066	0.7812	64	0.5564	6.4477	1.3234	5.4208
9	11.3295	8.4662	10.7841	0.8735	65	0.5349	6.4346	1.3017	5.4999
10	10.0285	8.3688	9.6496	0.9650	66	0.5142	6.4214	1.2807	5.5788
11	8.9673	8.2775	8.7244	1.0557	67	0.4944	6.4082	1.2603	5.6575
12	8.0857	8.1917	7.9558	1.1456	68	0.4753	6.3951	1.2405	5.7360
13	7.3419	8.1109	7.3075	1.2350	69	0.4570	6.3820	1.2213	5.8144
14	6.7063	8.0345	6.7535	1.3237	70	0.4395	6.3690	1.2026	5.8927
15	6.1572	7.9622	6.2749	1.4118	71	0.4226	6.3561	1.1844	5.9708
16	5.6782	7.8937	5.8573	1.4995	72	0.4064	6.3432	1.1668	6.0488
17	5.2568	7.8285	5.4900	1.5866	73	0.3908	6.3305	1.1497	6.1266
18	4.8834	7.7664	5.1644	1.6733	74	0.3758	6.3179	1.1330	6.2043
19	4.5503	7.7071	4.8738	1.7595	75	0.3614	6.3053	1.1168	6.2820
20	4.2515	7.6505	4.6131	1.8452	76	0.3476	6.2929	1.1010	6.3595
21	3.9820	7.5963	4.3778	1.9306	77	0.3343	6.2807	1.0857	6.4368
22	3.7377	7.5443	4.1645	2.0156	78	0.3215	6.2685	1.0707	6.5141
23	3.5154	7.4944	3.9703	2.1003	79	0.3092	6.2565	1.0561	6.5913
24	3.3122	7.4464	3.7926	2.1846	80	0.2974	6.2447	1.0419	6.6684
25	3.1259	7.4002	3.6296	2.2685	81	0.2860	6.2329	1.0281	6.7455

TABLE 12e (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
26	2.9545	7.3557	3.4795	2.3522	82	0.2751	6.2214	1.0146	6.8224
27	2.7962	7.3128	3.3408	2.4355	83	0.2646	6.2100	1.0015	6.8993
28	2.6497	7.2713	3.2124	2.5185	84	0.2545	6.1987	0.9887	6.9760
29	2.5138	7.2312	3.0930	2.6012	85	0.2448	6.1876	0.9762	7.0528
30	2.3873	7.1923	2.9819	2.6837	86	0.2355	6.1766	0.9640	7.1294
31	2.2693	7.1547	2.8781	2.7659	87	0.2265	6.1658	0.9520	7.2060
32	2.1591	7.1182	2.7811	2.8478	88	0.2179	6.1552	0.9404	7.2825
33	2.0558	7.0827	2.6901	2.9295	89	0.2096	6.1447	0.9291	7.3590
34	1.9590	7.0483	2.6046	3.0109	90	0.2016	6.1344	0.9180	7.4355
35	1.8614	7.0124	2.5237	3.0916	91	0.1939	6.1242	0.9071	7.5118
36	1.7777	6.9805	2.4480	3.1727	92	0.1865	6.1142	0.8965	7.5882
37	1.6987	6.9494	2.3766	3.2535	93	0.1794	6.1044	0.8862	7.6645
38	1.6241	6.9191	2.3090	3.3342	94	0.1726	6.0947	0.8760	7.7408
39	1.5535	6.8896	2.2450	3.4146	95	0.1660	6.0852	0.8661	7.8170
40	1.4867	6.8609	2.1843	3.4948	96	0.1597	6.0758	0.8565	7.8932
41	1.4233	6.8328	2.1266	3.5749	97	0.1536	6.0666	0.8470	7.9693
42	1.3632	6.8053	2.0718	3.6547	98	0.1478	6.0575	0.8377	8.0455
43	1.3060	6.7785	2.0197	3.7343	99	0.1422	6.0486	0.8286	8.1216
44	1.2517	6.7523	1.9699	3.8138	100	0.1368	6.0398	0.8198	8.1977
45	1.1999	6.7266	1.9225	3.8931	101	0.1316	6.0312	0.8111	8.2737
46	1.1507	6.7015	1.8772	3.9721	102	0.1266	6.0227	0.8026	8.3498
47	1.1037	6.6770	1.8339	4.0510	103	0.1217	6.0143	0.7942	8.4258
48	1.0588	6.6529	1.7924	4.1298	104	0.1171	6.0062	0.7860	8.5018
49	1.0161	6.6293	1.7527	4.2083	105	0.1127	5.9981	0.7780	8.5778
50	0.9752	6.6149	1.7158	4.2895	106	0.1084	5.9902	0.7702	8.6538
51	0.9361	6.6054	1.6810	4.3722	107	0.1043	5.9824	0.7625	8.7298
52	0.8988	6.5953	1.6474	4.4545	108	0.1003	5.9748	0.7549	8.8057
53	0.8630	6.5846	1.6150	4.5365	109	0.0965	5.9673	0.7476	8.8817
54	0.8289	6.5735	1.5837	4.6182	110	0.0928	5.9600	0.7403	8.9576
55	0.7961	6.5619	1.5536	4.6996	111	0.0893	5.9527	0.7332	9.0336
56	0.7648	6.5501	1.5245	4.7807	112	0.0859	5.9456	0.7262	9.1095
113	0.0826	5.9387	0.7194	9.1855	132	0.0396	5.8280	0.6101	10.6295
114	0.0795	5.9318	0.7126	9.2614	133	0.0381	5.8232	0.6052	10.7056
115	0.0765	5.9251	0.7060	9.3374	134	0.0366	5.8185	0.6005	10.7818
116	0.0736	5.9185	0.6996	9.4133	135	0.0352	5.8139	0.5958	10.8579
117	0.0708	5.9120	0.6932	9.4893	136	0.0339	5.8093	0.5912	10.9341
118	0.0681	5.9057	0.6870	9.5652	137	0.0326	5.8049	0.5866	11.0102
119	0.0655	5.8994	0.6808	9.6412	138	0.0314	5.8005	0.5821	11.0864
120	0.0630	5.8933	0.6748	9.7172	139	0.0302	5.7962	0.5777	11.1626
121	0.0606	5.8873	0.6689	9.7931	140	0.0290	5.7920	0.5734	11.2389
122	0.0583	5.8814	0.6631	9.8691	141	0.0279	5.7879	0.5691	11.3151
123	0.0561	5.8756	0.6574	9.9451	142	0.0269	5.7838	0.5649	11.3914
124	0.0540	5.8699	0.6517	10.0211	143	0.0258	5.7799	0.5608	11.4677
125	0.0519	5.8643	0.6462	10.0971	144	0.0249	5.7760	0.5567	11.5440
126	0.0499	5.8588	0.6408	10.1731	145	0.0239	5.7722	0.5527	11.6203
127	0.0480	5.8534	0.6354	10.2492	146	0.0230	5.7684	0.5487	11.6966
128	0.0462	5.8482	0.6302	10.3252	147	0.0221	5.7647	0.5448	11.7730
129	0.0445	5.8430	0.6250	10.4013	148	0.0213	5.7611	0.5410	11.8494
130	0.0428	5.8379	0.6200	10.4773	149	0.0205	5.7576	0.5372	11.9258
131	0.0411	5.8329	0.6150	10.5534	150	0.0197	5.7541	0.5334	12.0022

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12f Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er = 11.6

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	107.9253	11.3278	112.2672	0.1123	26	2.6229	8.5446	3.7502	2.5351
2	52.8209	11.0843	55.5270	0.2221	27	2.4782	8.4928	3.6003	2.6246
3	34.5132	10.8654	36.6506	0.3299	28	2.3444	8.4426	3.4615	2.7138
4	25.3888	10.6674	27.2364	0.4358	29	2.2202	8.3942	3.3325	2.8026
5	19.9316	10.4873	21.6044	0.5401	30	2.1047	8.3473	3.2124	2.8912
6	16.3052	10.3226	17.8617	0.6430	31	1.9970	8.3019	3.1003	2.9794
7	13.7232	10.1712	15.1973	0.7447	32	1.8884	8.2543	2.9948	3.0667
8	11.7929	10.0313	13.2059	0.8452	33	1.7964	8.2125	2.8967	3.1545
9	10.2964	9.9016	11.6625	0.9447	34	1.7101	8.1718	2.8045	3.2420
10	9.1031	9.7809	10.4321	1.0432	35	1.6291	8.1324	2.7178	3.3293
11	8.1299	9.6682	9.4289	1.1409	36	1.5527	8.0939	2.6361	3.4163
12	7.3215	9.5625	8.5958	1.2378	37	1.4807	8.0565	2.5589	3.5031
13	6.6397	9.4632	7.8932	1.3340	38	1.4128	8.0201	2.4859	3.5897
14	6.0573	9.3696	7.2931	1.4294	39	1.3485	7.9846	2.4168	3.6759
15	5.5541	9.2812	6.7747	1.5243	40	1.2877	7.9499	2.3513	3.7620
16	5.1153	9.1975	6.3226	1.6186	41	1.2301	7.9161	2.2890	3.8478
17	4.7294	9.1180	5.9249	1.7123	42	1.1754	7.8830	2.2299	3.9335
18	4.3875	9.0424	5.5725	1.8055	43	1.1235	7.8507	2.1735	4.0189
19	4.0826	8.9703	5.2581	1.8982	44	1.0742	7.8192	2.1199	4.1041
20	3.8091	8.9016	4.9760	1.9904	45	1.0273	7.7883	2.0687	4.1890
21	3.5625	8.8358	4.7215	2.0822	46	0.9827	7.7654	2.0207	4.2758
22	3.3390	8.7728	4.4908	2.1736	47	0.9402	7.7532	1.9762	4.3653
23	3.1357	8.7124	4.2807	2.2645	48	0.8998	7.7401	1.9334	4.4545
24	2.9499	8.6543	4.0887	2.3551	49	0.8612	7.7262	1.8922	4.5432
25	2.7796	8.5984	3.9125	2.4453	50	0.8244	7.7117	1.8526	4.6315
51	0.7893	7.6966	1.8145	4.7195	101	0.0957	6.9814	0.8726	8.9017
52	0.7558	7.6810	1.7778	4.8072	102	0.0917	6.9719	0.8635	8.9837
53	0.7238	7.6650	1.7425	4.8945	103	0.0880	6.9625	0.8545	9.0657
54	0.6933	7.6487	1.7084	4.9816	104	0.0844	6.9534	0.8458	9.1477
55	0.6641	7.6322	1.6755	5.0683	105	0.0809	6.9444	0.8372	9.2297
56	0.6362	7.6154	1.6438	5.1548	106	0.0776	6.9356	0.8287	9.3116
57	0.6095	7.5985	1.6131	5.2410	107	0.0744	6.9269	0.8205	9.3936
58	0.5840	7.5815	1.5835	5.3270	108	0.0714	6.9185	0.8124	9.4756
59	0.5596	7.5643	1.5549	5.4127	109	0.0684	6.9102	0.8044	9.5576
60	0.5363	7.5472	1.5273	5.4982	110	0.0656	6.9020	0.7967	9.6396
61	0.5139	7.5301	1.5005	5.5835	111	0.0630	6.8940	0.7890	9.7216
62	0.4925	7.5129	1.4747	5.6686	112	0.0604	6.8862	0.7815	9.8036
63	0.4721	7.4959	1.4496	5.7535	113	0.0579	6.8786	0.7742	9.8857
64	0.4525	7.4789	1.4253	5.8382	114	0.0555	6.8710	0.7670	9.9677
65	0.4337	7.4620	1.4018	5.9227	115	0.0533	6.8637	0.7599	10.0498
66	0.4158	7.4452	1.3790	6.0071	116	0.0511	6.8565	0.7530	10.1318
67	0.3986	7.4286	1.3569	6.0913	117	0.0490	6.8494	0.7461	10.2139
68	0.3821	7.4121	1.3355	6.1753	118	0.0470	6.8425	0.7394	10.2960
69	0.3663	7.3957	1.3147	6.2592	119	0.0450	6.8357	0.7329	10.3781
70	0.3512	7.3795	1.2945	6.3430	120	0.0432	6.8290	0.7264	10.4602
71	0.3367	7.3635	1.2749	6.4266	121	0.0414	6.8225	0.7201	10.5423
72	0.3228	7.3477	1.2558	6.5101	122	0.0397	6.8161	0.7138	10.6245
73	0.3095	7.3321	1.2373	6.5935	123	0.0381	6.8099	0.7077	10.7067
74	0.2967	7.3166	1.2193	6.6768	124	0.0365	6.8037	0.7017	10.7888
75	0.2845	7.3014	1.2018	6.7600	125	0.0350	6.7977	0.6957	10.8711
76	0.2728	7.2864	1.1847	6.8430	126	0.0336	6.7919	0.6899	10.9533
77	0.2616	7.2716	1.1682	6.9260	127	0.0322	6.7861	0.6842	11.0355
78	0.2508	7.2570	1.1520	7.0089	128	0.0309	6.7804	0.6786	11.1178
79	0.2405	7.2426	1.1363	7.0918	129	0.0296	6.7749	0.6730	11.2001
80	0.2306	7.2285	1.1210	7.1745	130	0.0284	6.7695	0.6676	11.2824

TABLE 12f (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
81	0.2211	7.2146	1.1061	7.2572	131	0.0273	6.7642	0.6622	11.3647
82	0.2121	7.2009	1.0916	7.3398	132	0.0261	6.7590	0.6570	11.4471
83	0.2033	7.1874	1.0774	7.4224	133	0.0251	6.7539	0.6518	11.5294
84	0.1950	7.1741	1.0636	7.5049	134	0.0240	6.7489	0.6467	11.6118
85	0.1870	7.1611	1.0501	7.5873	135	0.0231	6.7440	0.6417	11.6943
86	0.1793	7.1483	1.0370	7.6697	136	0.0221	6.7392	0.6367	11.7767
87	0.1719	7.1357	1.0242	7.7520	137	0.0212	6.7345	0.6318	11.8592
88	0.1649	7.1233	1.0117	7.8343	138	0.0203	6.7300	0.6271	11.9417
89	0.1581	7.1111	0.9994	7.9166	139	0.0195	6.7255	0.6223	12.0242
90	0.1516	7.0992	0.9875	7.9988	140	0.0187	6.7210	0.6177	12.1067
91	0.1454	7.0874	0.9758	8.0810	141	0.0179	6.7167	0.6131	12.1893
92	0.1394	7.0759	0.9645	8.1632	142	0.0172	6.7125	0.6086	12.2718
93	0.1337	7.0646	0.9533	8.2453	143	0.0165	6.7084	0.6042	12.3545
94	0.1282	7.0535	0.9424	8.3274	144	0.0158	6.7043	0.5998	12.4371
95	0.1230	7.0426	0.9318	8.4095	145	0.0152	6.7003	0.5955	12.5197
96	0.1179	7.0319	0.9214	8.4916	146	0.0146	6.6964	0.5912	12.6024
97	0.1131	7.0214	0.9112	8.5736	147	0.0140	6.6926	0.5870	12.6851
98	0.1085	7.0111	0.9013	8.6556	148	0.0134	6.6889	0.5829	12.7679
99	0.1040	7.0010	0.8915	8.7377	149	0.0128	6.6852	0.5788	12.8506
100	0.0998	6.9911	0.8820	8.8197	150	0.0123	6.6817	0.5748	12.9334

Notes: Calculation of W/H has an error of less than 1%.
Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12g Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er =11.9

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	106.5298	11.6168	113.6899	0.1137	55	0.6453	7.8084	1.6947	5.1265
2	52.1270	11.3637	56.2223	0.2249	56	0.6179	7.7908	1.6626	5.2139
3	34.0530	11.1365	37.1049	0.3339	57	0.5918	7.7731	1.6316	5.3009
4	25.0454	10.9312	27.5710	0.4411	58	0.5667	7.7554	1.6016	5.3878
5	19.6583	10.7446	21.8678	0.5467	59	0.5428	7.7375	1.5726	5.4744
6	16.0785	10.5741	18.0780	0.6508	60	0.5199	7.7197	1.5446	5.5607
7	13.5298	10.4175	15.3802	0.7536	61	0.4980	7.7019	1.5176	5.6469
8	11.6245	10.2730	13.3640	0.8553	62	0.4771	7.6841	1.4914	5.7328
9	10.1474	10.1390	11.8014	0.9559	63	0.4571	7.6664	1.4660	5.8186
10	8.9696	10.0144	10.5558	1.0556	64	0.4379	7.6488	1.4414	5.9041
11	8.0091	9.8981	9.5403	1.1544	65	0.4195	7.6313	1.4176	5.9895
12	7.2113	9.7891	8.6970	1.2524	66	0.4020	7.6140	1.3946	6.0748
13	6.5385	9.6867	7.9859	1.3496	67	0.3852	7.5968	1.3722	6.1598
14	5.9637	9.5902	7.3784	1.4462	68	0.3691	7.5797	1.3505	6.2447
15	5.4672	9.4991	6.8538	1.5421	69	0.3536	7.5628	1.3295	6.3295
16	5.0342	9.4128	6.3961	1.6374	70	0.3389	7.5462	1.3090	6.4142
17	4.6534	9.3309	5.9937	1.7322	71	0.3247	7.5297	1.2892	6.4987
18	4.3160	9.2531	5.6370	1.8264	72	0.3112	7.5134	1.2699	6.5831
19	4.0152	9.1789	5.3189	1.9201	73	0.2982	7.4973	1.2511	6.6674
20	3.7454	9.1081	5.0334	2.0134	74	0.2858	7.4814	1.2329	6.7516
21	3.5020	9.0404	4.7759	2.1062	75	0.2739	7.4658	1.2152	6.8356
22	3.2816	8.9755	4.5424	2.1985	76	0.2625	7.4504	1.1980	6.9196
23	3.0810	8.9134	4.3298	2.2905	77	0.2516	7.4352	1.1812	7.0035
24	2.8977	8.8536	4.1355	2.3820	78	0.2411	7.4202	1.1649	7.0873
25	2.7297	8.7961	3.9572	2.4732	79	0.2311	7.4055	1.1490	7.1711

TABLE 12g (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
26	2.5751	8.7408	3.7930	2.5641	80	0.2215	7.3910	1.1335	7.2547
27	2.4324	8.6874	3.6413	2.6545	81	0.2123	7.3767	1.1185	7.3383
28	2.3004	8.6359	3.5009	2.7447	82	0.2035	7.3627	1.1038	7.4218
29	2.1779	8.5861	3.3704	2.8345	83	0.1950	7.3489	1.0895	7.5053
30	2.0640	8.5378	3.2489	2.9240	84	0.1869	7.3353	1.0755	7.5887
31	1.9578	8.4911	3.1355	3.0132	85	0.1791	7.3220	1.0619	7.6721
32	1.8513	8.4425	3.0288	3.1014	86	0.1717	7.3089	1.0486	7.7554
33	1.7607	8.3995	2.9295	3.1902	87	0.1646	7.2960	1.0356	7.8386
34	1.6756	8.3577	2.8362	3.2787	88	0.1577	7.2834	1.0230	7.9219
35	1.5956	8.3171	2.7485	3.3669	89	0.1512	7.2709	1.0106	8.0051
36	1.5204	8.2776	2.6658	3.4549	90	0.1449	7.2588	0.9985	8.0882
37	1.4494	8.2391	2.5877	3.5426	91	0.1389	7.2468	0.9868	8.1713
38	1.3824	8.2016	2.5139	3.6301	92	0.1331	7.2350	0.9752	8.2544
39	1.3190	8.1651	2.4440	3.7173	93	0.1276	7.2235	0.9640	8.3375
40	1.2591	8.1295	2.3777	3.8043	94	0.1223	7.2122	0.9530	8.4206
41	1.2023	8.0947	2.3147	3.8910	95	0.1172	7.2011	0.9422	8.5036
42	1.1484	8.0607	2.2548	3.9775	96	0.1124	7.1902	0.9317	8.5866
43	1.0973	8.0275	2.1979	4.0639	97	0.1077	7.1795	0.9214	8.6696
44	1.0487	7.9951	2.1436	4.1499	98	0.1033	7.1691	0.9113	8.7526
45	1.0026	7.9633	2.0918	4.2358	99	0.0990	7.1588	0.9015	8.8356
46	0.9586	7.9500	2.0446	4.3263	100	0.0949	7.1487	0.8919	8.9185
47	0.9168	7.9367	1.9994	4.4167	101	0.0909	7.1388	0.8824	9.0015
48	0.8770	7.9226	1.9560	4.5067	102	0.0872	7.1292	0.8732	9.0845
49	0.8391	7.9078	1.9143	4.5962	103	0.0836	7.1197	0.8641	9.1674
50	0.8029	7.8923	1.8742	4.6854	104	0.0801	7.1104	0.8552	9.2504
51	0.7684	7.8762	1.8356	4.7743	105	0.0768	7.1012	0.8466	9.3333
52	0.7354	7.8598	1.7984	4.8628	106	0.0736	7.0923	0.8380	9.4163
53	0.7040	7.8429	1.7626	4.9510	107	0.0705	7.0835	0.8297	9.4992
54	0.6740	7.8258	1.7280	5.0389	108	0.0676	7.0749	0.8215	9.5822
109	0.0648	7.0665	0.8135	9.6652	130	0.0266	6.9243	0.6752	11.4107
110	0.0621	7.0583	0.8056	9.7481	131	0.0255	6.9190	0.6698	11.4940
111	0.0595	7.0502	0.7979	9.8311	132	0.0245	6.9138	0.6645	11.5774
112	0.0571	7.0423	0.7903	9.9141	133	0.0235	6.9086	0.6592	11.6608
113	0.0547	7.0345	0.7829	9.9971	134	0.0225	6.9036	0.6541	11.7442
114	0.0524	7.0269	0.7756	10.0801	135	0.0216	6.8987	0.6490	11.8276
115	0.0503	7.0195	0.7685	10.1632	136	0.0207	6.8939	0.6440	11.9110
116	0.0482	7.0122	0.7615	10.2462	137	0.0198	6.8892	0.6391	11.9945
117	0.0462	7.0050	0.7546	10.3293	138	0.0190	6.8845	0.6342	12.0780
118	0.0443	6.9980	0.7478	10.4123	139	0.0182	6.8800	0.6294	12.1615
119	0.0424	6.9911	0.7412	10.4954	140	0.0174	6.8756	0.6247	12.2451
120	0.0407	6.9844	0.7346	10.5785	141	0.0167	6.8713	0.6201	12.3287
121	0.0390	6.9778	0.7282	10.6617	142	0.0160	6.8670	0.6156	12.4123
122	0.0374	6.9714	0.7219	10.7448	143	0.0154	6.8628	0.6111	12.4959
123	0.0358	6.9651	0.7157	10.8280	144	0.0147	6.8588	0.6067	12.5795
124	0.0343	6.9589	0.7096	10.9112	145	0.0141	6.8548	0.6023	12.6632
125	0.0329	6.9528	0.7036	10.9944	146	0.0135	6.8509	0.5980	12.7469
126	0.0316	6.9469	0.6978	11.0776	147	0.0130	6.8471	0.5938	12.8306
127	0.0302	6.9411	0.6920	11.1608	148	0.0124	6.8433	0.5896	12.9144
128	0.0290	6.9354	0.6863	11.2441	149	0.0119	6.8396	0.5855	12.9982
129	0.0278	6.9298	0.6807	11.3274	150	0.0114	6.8361	0.5814	13.0820

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12h Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er =12.88

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	102.3159	12.5596	118.2135	0.1182	24	2.7402	9.5011	4.2841	2.4676
2	50.0315	12.2746	58.4323	0.2337	25	2.5791	9.4384	4.0991	2.5619
3	32.6633	12.0197	38.5483	0.3469	26	2.4309	9.3779	3.9288	2.6559
4	24.0084	11.7903	28.6339	0.4581	27	2.2942	9.3197	3.7715	2.7494
5	18.8329	11.5823	22.7043	0.5676	28	2.1677	9.2635	3.6258	2.8427
6	15.3941	11.3928	18.7648	0.6755	29	2.0503	9.2091	3.4905	2.9355
7	12.9460	11.2191	15.9610	0.7821	30	1.9316	9.1518	3.3637	3.0273
8	11.1161	11.0591	13.8660	0.8874	31	1.8324	9.1020	3.2463	3.1197
9	9.6977	10.9111	12.2425	0.9916	32	1.7397	9.0536	3.1365	3.2117
10	8.5668	10.7736	10.9487	1.0949	33	1.6528	9.0068	3.0335	3.3035
11	7.6446	10.6455	9.8939	1.1972	34	1.5714	8.9612	2.9369	3.3950
12	6.8787	10.5255	9.0182	1.2986	35	1.4948	8.9170	2.8459	3.4862
13	6.2329	10.4130	8.2799	1.3993	36	1.4228	8.8739	2.7602	3.5772
14	5.6812	10.3071	7.6492	1.4993	37	1.3549	8.8320	2.6792	3.6678
15	5.2048	10.2071	7.1046	1.5985	38	1.2908	8.7911	2.6027	3.7582
16	4.7893	10.1125	6.6296	1.6972	39	1.2302	8.7513	2.5302	3.8484
17	4.4239	10.0228	6.2119	1.7952	40	1.1729	8.7125	2.4614	3.9383
18	4.1003	9.9376	5.8418	1.8927	41	1.1186	8.6745	2.3962	4.0280
19	3.8117	9.8565	5.5117	1.9897	42	1.0672	8.6375	2.3341	4.1174
20	3.5529	9.7791	5.2155	2.0862	43	1.0184	8.6014	2.2751	4.2066
21	3.3196	9.7051	4.9483	2.1822	44	0.9721	8.5791	2.2205	4.2989
22	3.1082	9.6342	4.7061	2.2778	45	0.9281	8.5645	2.1693	4.3928
23	2.9159	9.5663	4.4856	2.3729	46	0.8863	8.5487	2.1202	4.4863
47	0.8465	8.5321	2.0731	4.5794	99	0.0845	7.6726	0.9333	9.1471
48	0.8086	8.5147	2.0278	4.6720	100	0.0808	7.6620	0.9233	9.2331
49	0.7726	8.4967	1.9843	4.7643	101	0.0774	7.6516	0.9136	9.3192
50	0.7383	8.4781	1.9425	4.8562	102	0.0740	7.6415	0.9040	9.4052
51	0.7056	8.4591	1.9023	4.9478	103	0.0709	7.6315	0.8946	9.4912
52	0.6744	8.4397	1.8635	5.0390	104	0.0678	7.6218	0.8855	9.5773
53	0.6447	8.4201	1.8263	5.1299	105	0.0649	7.6123	0.8765	9.6633
54	0.6163	8.4002	1.7903	5.2206	106	0.0621	7.6030	0.8677	9.7494
55	0.5893	8.3802	1.7557	5.3109	107	0.0595	7.5938	0.8591	9.8354
56	0.5634	8.3601	1.7223	5.4010	108	0.0569	7.5849	0.8506	9.9215
57	0.5388	8.3399	1.6900	5.4908	109	0.0545	7.5761	0.8423	10.0076
58	0.5152	8.3198	1.6588	5.5804	110	0.0521	7.5675	0.8342	10.0937
59	0.4928	8.2996	1.6288	5.6697	111	0.0499	7.5592	0.8262	10.1798
60	0.4713	8.2796	1.5997	5.7588	112	0.0477	7.5509	0.8184	10.2659
61	0.4508	8.2596	1.5716	5.8477	113	0.0457	7.5429	0.8107	10.3521
62	0.4312	8.2397	1.5443	5.9365	114	0.0437	7.5350	0.8032	10.4382
63	0.4125	8.2200	1.5180	6.0250	115	0.0418	7.5273	0.7958	10.5244
64	0.3946	8.2005	1.4925	6.1133	116	0.0400	7.5198	0.7885	10.6106
65	0.3775	8.1811	1.4678	6.2015	117	0.0383	7.5124	0.7814	10.6968
66	0.3611	8.1619	1.4439	6.2895	118	0.0367	7.5052	0.7744	10.7830
67	0.3455	8.1429	1.4207	6.3774	119	0.0351	7.4981	0.7676	10.8693
68	0.3306	8.1242	1.3982	6.4652	120	0.0336	7.4912	0.7608	10.9556
69	0.3163	8.1057	1.3763	6.5528	121	0.0321	7.4844	0.7542	11.0419
70	0.3026	8.0874	1.3551	6.6402	122	0.0308	7.4778	0.7477	11.1282
71	0.2895	8.0694	1.3346	6.7276	123	0.0294	7.4713	0.7413	11.2146
72	0.2770	8.0517	1.3146	6.8148	124	0.0282	7.4649	0.7350	11.3009
73	0.2651	8.0342	1.2952	6.9020	125	0.0270	7.4587	0.7288	11.3873
74	0.2536	8.0169	1.2763	6.9890	126	0.0258	7.4526	0.7227	11.4737
75	0.2427	8.0000	1.2579	7.0759	127	0.0247	7.4467	0.7167	11.5602
76	0.2323	7.9833	1.2401	7.1628	128	0.0236	7.4409	0.7109	11.6466
77	0.2222	7.9669	1.2227	7.2496	129	0.0226	7.4352	0.7051	11.7331
78	0.2127	7.9507	1.2058	7.3363	130	0.0217	7.4296	0.6994	11.8197
79	0.2035	7.9348	1.1894	7.4229	131	0.0207	7.4241	0.6938	11.9062

TABLE 12h (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
80	0.1947	7.9192	1.1734	7.5095	132	0.0198	7.4188	0.6883	11.9928
81	0.1864	7.9039	1.1578	7.5960	133	0.0190	7.4136	0.6829	12.0794
82	0.1783	7.8889	1.1425	7.6825	134	0.0182	7.4085	0.6775	12.1660
83	0.1707	7.8741	1.1277	7.7689	135	0.0174	7.4035	0.6723	12.2527
84	0.1633	7.8596	1.1133	7.8552	136	0.0166	7.3986	0.6671	12.3393
85	0.1563	7.8453	1.0992	7.9415	137	0.0159	7.3938	0.6621	12.4261
86	0.1496	7.8314	1.0854	8.0278	138	0.0152	7.3891	0.6570	12.5128
87	0.1431	7.8176	1.0720	8.1140	139	0.0146	7.3845	0.6521	12.5996
88	0.1370	7.8042	1.0589	8.2002	140	0.0140	7.3800	0.6473	12.6863
89	0.1311	7.7910	1.0461	8.2864	141	0.0134	7.3756	0.6425	12.7732
90	0.1255	7.7781	1.0336	8.3725	142	0.0128	7.3713	0.6378	12.8600
91	0.1201	7.7654	1.0215	8.4587	143	0.0122	7.3671	0.6331	12.9469
92	0.1149	7.7529	1.0095	8.5448	144	0.0117	7.3630	0.6286	13.0338
93	0.1100	7.7407	0.9979	8.6309	145	0.0112	7.3590	0.6241	13.1207
94	0.1052	7.7288	0.9865	8.7169	146	0.0107	7.3551	0.6196	13.2077
95	0.1007	7.7171	0.9754	8.8030	147	0.0103	7.3512	0.6152	13.2946
96	0.0964	7.7056	0.9645	8.8890	148	0.0098	7.3475	0.6109	13.3817
97	0.0922	7.6944	0.9539	8.9751	149	0.0094	7.3438	0.6067	13.4687
98	0.0883	7.6833	0.9435	9.0611	150	0.0090	7.3402	0.6025	13.5558

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12i Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er =35

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	61.2525	33.5453	193.1947	0.1932	57	0.1141	20.3209	2.6380	8.5709
2	29.6179	32.3412	94.8478	0.3794	58	0.1063	20.2652	2.5890	8.7093
3	19.1302	31.3265	62.2321	0.5601	59	0.0990	20.2113	2.5417	8.8477
4	13.9141	30.4569	46.0216	0.7363	60	0.0923	20.1591	2.4961	8.9860
5	10.8011	29.7005	36.3573	0.9089	61	0.0860	20.1087	2.4521	9.1243
6	8.7367	29.0345	29.9561	1.0784	62	0.0801	20.0598	2.4096	9.2626
7	7.2699	28.4417	25.4132	1.2452	63	0.0746	20.0126	2.3686	9.4010
8	6.1757	27.9094	22.0275	1.4098	64	0.0695	19.9670	2.3289	9.5393
9	5.3291	27.4273	19.4101	1.5722	65	0.0648	19.9228	2.2906	9.6776
10	4.6555	26.9878	17.3286	1.7329	66	0.0604	19.8801	2.2534	9.8160
11	4.1073	26.5845	15.6351	1.8918	67	0.0562	19.8389	2.2175	9.9543
12	3.6529	26.2124	14.2315	2.0493	68	0.0524	19.7990	2.1827	10.0928
13	3.2705	25.8674	13.0501	2.2055	69	0.0488	19.7604	2.1490	10.2312
14	2.9445	25.5460	12.0424	2.3603	70	0.0455	19.7232	2.1163	10.3697
15	2.6635	25.2453	11.1732	2.5140	71	0.0424	19.6872	2.0846	10.5082
16	2.4189	24.9629	10.4161	2.6665	72	0.0395	19.6525	2.0538	10.6468
17	2.2042	24.6968	9.7510	2.8180	73	0.0368	19.6189	2.0239	10.7855
18	2.0144	24.4452	9.1623	2.9686	74	0.0343	19.5865	1.9949	10.9242
19	1.8358	24.1924	8.6351	3.1173	75	0.0319	19.5551	1.9667	11.0630
20	1.6891	23.9716	8.1658	3.2663	76	0.0298	19.5249	1.9394	11.2018
21	1.5569	23.7610	7.7427	3.4145	77	0.0277	19.4957	1.9127	11.3407
22	1.4372	23.5597	7.3594	3.5619	78	0.0258	19.4674	1.8869	11.4796
23	1.3284	23.3669	7.0106	3.7086	79	0.0241	19.4402	1.8617	11.6187
24	1.2292	23.1820	6.6918	3.8545	80	0.0224	19.4139	1.8372	11.7578
25	1.1385	23.0043	6.3995	3.9997	81	0.0209	19.3884	1.8133	11.8970

TABLE 12i (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
26	1.0554	22.8334	6.1304	4.1442	82	0.0195	19.3639	1.7900	12.0362
27	0.9790	22.6972	5.8857	4.2907	83	0.0181	19.3402	1.7674	12.1755
28	0.9087	22.6289	5.6670	4.4429	84	0.0169	19.3173	1.7453	12.3149
29	0.8438	22.5532	5.4624	4.5939	85	0.0157	19.2952	1.7238	12.4544
30	0.7840	22.4719	5.2708	4.7437	86	0.0147	19.2739	1.7028	12.5940
31	0.7286	22.3863	5.0911	4.8925	87	0.0137	19.2533	1.6823	12.7336
32	0.6774	22.2977	4.9222	5.0403	88	0.0127	19.2334	1.6624	12.8733
33	0.6300	22.2069	4.7633	5.1873	89	0.0119	19.2142	1.6429	13.0131
34	0.5861	22.1148	4.6136	5.3333	90	0.0111	19.1957	1.6238	13.1530
35	0.5453	22.0220	4.4724	5.4787	91	0.0103	19.1778	1.6052	13.2929
36	0.5074	21.9291	4.3390	5.6233	92	0.0096	19.1605	1.5871	13.4329
37	0.4723	21.8365	4.2128	5.7673	93	0.0089	19.1439	1.5693	13.5730
38	0.4397	21.7445	4.0933	5.9107	94	0.0083	19.1278	1.5520	13.7132
39	0.4093	21.6536	3.9800	6.0535	95	0.0078	19.1122	1.5350	13.8535
40	0.3811	21.5638	3.8724	6.1959	96	0.0072	19.0972	1.5184	13.9938
41	0.3549	21.4755	3.7702	6.3378	97	0.0067	19.0828	1.5022	14.1342
42	0.3305	21.3889	3.6730	6.4792	98	0.0063	19.0688	1.4863	14.2747
43	0.3078	21.3039	3.5805	6.6203	99	0.0059	19.0553	1.4708	14.4152
44	0.2867	21.2208	3.4923	6.7610	100	0.0055	19.0422	1.4556	14.5559
45	0.2670	21.1396	3.4081	6.9014	101	0.0051	19.0297	1.4407	14.6966
46	0.2487	21.0603	3.3278	7.0416	102	0.0047	19.0175	1.4261	14.8374
47	0.2317	20.9831	3.2510	7.1814	103	0.0044	19.0058	1.4118	14.9782
48	0.2158	20.9078	3.1776	7.3211	104	0.0041	18.9945	1.3978	15.1191
49	0.2010	20.8346	3.1073	7.4605	105	0.0038	18.9836	1.3841	15.2601
50	0.1873	20.7635	3.0399	7.5998	106	0.0036	18.9730	1.3707	15.4012
51	0.1745	20.6944	2.9753	7.7388	107	0.0033	18.9629	1.3575	15.5423
52	0.1625	20.6272	2.9134	7.8778	108	0.0031	18.9530	1.3446	15.6835
53	0.1514	20.5621	2.8539	8.0166	109	0.0029	18.9436	1.3319	15.8247
54	0.1411	20.4989	2.7967	8.1553	110	0.0027	18.9344	1.3195	15.9661
55	0.1314	20.4377	2.7418	8.2939	111	0.0025	18.9256	1.3073	16.1074
56	0.1224	20.3784	2.6889	8.4324	112	0.0023	18.9170	1.2954	16.2489
113	0.0022	18.9088	1.2836	16.3904	132	0.0006	18.7969	1.0956	19.0896
114	0.0020	18.9009	1.2721	16.5320	133	0.0005	18.7928	1.0872	19.2321
115	0.0019	18.8932	1.2608	16.6736	134	0.0005	18.7889	1.0790	19.3747
116	0.0018	18.8858	1.2497	16.8153	135	0.0005	18.7851	1.0709	19.5173
117	0.0016	18.8786	1.2387	16.9571	136	0.0004	18.7815	1.0629	19.6600
118	0.0015	18.8717	1.2280	17.0989	137	0.0004	18.7779	1.0551	19.8027
119	0.0014	18.8651	1.2175	17.2407	138	0.0004	18.7745	1.0473	19.9454
120	0.0013	18.8587	1.2071	17.3826	139	0.0003	18.7713	1.0397	20.0882
121	0.0012	18.8524	1.1970	17.5246	140	0.0003	18.7681	1.0322	20.2310
122	0.0012	18.8465	1.1870	17.6666	141	0.0003	18.7650	1.0248	20.3738
123	0.0011	18.8407	1.1771	17.8087	142	0.0003	18.7621	1.0175	20.5167
124	0.0010	18.8351	1.1675	17.9508	143	0.0003	18.7592	1.0103	20.6596
125	0.0009	18.8297	1.1580	18.0930	144	0.0002	18.7565	1.0032	20.8026
126	0.0009	18.8245	1.1486	18.2352	145	0.0002	18.7538	0.9962	20.9456
127	0.0008	18.8195	1.1394	18.3775	146	0.0002	18.7513	0.9893	21.0886
128	0.0008	18.8146	1.1304	18.5198	147	0.0002	18.7488	0.9825	21.2316
129	0.0007	18.8100	1.1215	18.6622	148	0.0002	18.7464	0.9758	21.3747
130	0.0007	18.8055	1.1127	18.8046	149	0.0002	18.7441	0.9692	21.5178
131	0.0006	18.8011	1.1041	18.9471	150	0.0002	18.7419	0.9627	21.6609

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.

TABLE 12j Zero Thickness Microstrip Dimensions, Effective Dielectric Constant, and PUL Capacitance and Inductance for Er =85

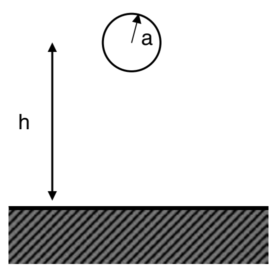
Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
1	38.5925	79.6824	297.7560	0.2978	26	0.3739	51.7462	9.2288	6.2387
2	18.3705	75.6651	145.0765	0.5803	27	0.3349	51.4124	8.8583	6.4577
3	11.6859	72.5009	94.6738	0.8521	28	0.3000	51.0886	8.5150	6.6757
4	8.3710	69.9235	69.7318	1.1157	29	0.2688	50.7757	8.1961	6.8930
5	6.3982	67.7680	54.9189	1.3730	30	0.2409	50.4745	7.8994	7.1095
6	5.0937	65.9270	45.1398	1.6250	31	0.2159	50.1850	7.6226	7.3253
7	4.1695	64.3277	38.2191	1.8727	32	0.1935	49.9076	7.3640	7.5407
8	3.4821	62.9184	33.0733	2.1167	33	0.1734	49.6422	7.1218	7.7557
9	2.9517	61.6613	29.1034	2.3574	34	0.1554	49.3886	6.8947	7.9702
10	2.5309	60.5285	25.9513	2.5951	35	0.1393	49.1467	6.6813	8.1845
11	2.1895	59.4981	23.3904	2.8302	36	0.1249	48.9161	6.4804	8.3986
12	1.8946	58.5089	21.2622	3.0618	37	0.1119	48.6966	6.2911	8.6125
13	1.6651	57.6611	19.4840	3.2928	38	0.1003	48.4877	6.1124	8.8263
14	1.4696	56.8730	17.9682	3.5218	39	0.0899	48.2891	5.9435	9.0400
15	1.3012	56.1365	16.6614	3.7488	40	0.0806	48.1003	5.7835	9.2537
16	1.1551	55.4455	15.5236	3.9740	41	0.0723	47.9210	5.6319	9.4673
17	1.0274	54.7949	14.5245	4.1976	42	0.0648	47.7509	5.4881	9.6810
18	0.9153	54.4535	13.6748	4.4306	43	0.0581	47.5893	5.3514	9.8947
19	0.8165	54.1607	12.9202	4.6642	44	0.0521	47.4361	5.2213	10.1085
20	0.7290	53.8383	12.2376	4.8950	45	0.0467	47.2907	5.0975	10.3224
21	0.6514	53.4976	11.6179	5.1235	46	0.0418	47.1529	4.9794	10.5364
22	0.5825	53.1467	11.0534	5.3498	47	0.0375	47.0222	4.8667	10.7505
23	0.5211	52.7921	10.5375	5.5743	48	0.0336	46.8984	4.7590	10.9648
24	0.4664	52.4383	10.0645	5.7972	49	0.0301	46.7810	4.6561	11.1792
25	0.4176	52.0888	9.6297	6.0185	50	0.0270	46.6697	4.5575	11.3938
51	0.0242	46.5643	4.4631	11.6085	101	0.0001	44.8028	2.2106	22.5503
52	0.0217	46.4644	4.3726	11.8234	102	0.0001	44.7963	2.1888	22.7720
53	0.0195	46.3698	4.2857	12.0385	103	0.0001	44.7901	2.1674	22.9936
54	0.0175	46.2801	4.2023	12.2538	104	0.0001	44.7842	2.1464	23.2153
55	0.0156	46.1952	4.1221	12.4693	105	0.0001	44.7787	2.1258	23.4371
56	0.0140	46.1148	4.0449	12.6849	106	0.0001	44.7734	2.1056	23.6589
57	0.0126	46.0386	3.9707	12.9008	107	0.0001	44.7685	2.0858	23.8808
58	0.0113	45.9665	3.8992	13.1168	108	0.0000	44.7638	2.0664	24.1027
59	0.0101	45.8982	3.8302	13.3330	109	0.0000	44.7593	2.0474	24.3247
60	0.0091	45.8335	3.7637	13.5495	110	0.0000	44.7551	2.0287	24.5467
61	0.0081	45.7722	3.6996	13.7661	111	0.0000	44.7511	2.0103	24.7688
62	0.0073	45.7142	3.6376	13.9829	112	0.0000	44.7473	1.9923	24.9908
63	0.0065	45.6592	3.5777	14.1999	113	0.0000	44.7437	1.9745	25.2130
64	0.0059	45.6072	3.5198	14.4170	114	0.0000	44.7404	1.9572	25.4351
65	0.0052	45.5579	3.4638	14.6344	115	0.0000	44.7371	1.9401	25.6573
66	0.0047	45.5112	3.4095	14.8519	116	0.0000	44.7341	1.9233	25.8795
67	0.0042	45.4670	3.3570	15.0696	117	0.0000	44.7312	1.9068	26.1018
68	0.0038	45.4252	3.3061	15.2875	118	0.0000	44.7285	1.8906	26.3241
69	0.0034	45.3856	3.2568	15.5056	119	0.0000	44.7259	1.8746	26.5464
70	0.0030	45.3481	3.2089	15.7238	120	0.0000	44.7235	1.8589	26.7688
71	0.0027	45.3126	3.1625	15.9422	121	0.0000	44.7212	1.8435	26.9911
72	0.0024	45.2789	3.1174	16.1607	122	0.0000	44.7190	1.8284	27.2135
73	0.0022	45.2471	3.0736	16.3794	123	0.0000	44.7169	1.8135	27.4360
74	0.0020	45.2169	3.0311	16.5982	124	0.0000	44.7149	1.7988	27.6584
75	0.0018	45.1884	2.9897	16.8172	125	0.0000	44.7131	1.7844	27.8809
76	0.0016	45.1614	2.9495	17.0363	126	0.0000	44.7113	1.7702	28.1034
77	0.0014	45.1358	2.9104	17.2556	127	0.0000	44.7097	1.7562	28.3259
78	0.0013	45.1115	2.8723	17.4750	128	0.0000	44.7081	1.7425	28.5484
79	0.0011	45.0886	2.8352	17.6946	129	0.0000	44.7066	1.7289	28.7710
80	0.0010	45.0669	2.7991	17.9142	130	0.0000	44.7052	1.7156	28.9936

TABLE 12j (continued)

Zo [Ohms]	W/h	Keff	pF/cm	nH/mm	Zo [Ohms]	W/h	Keff	pF/cm	nH/mm
81	0.0009	45.0463	2.7639	18.1340	131	0.0000	44.7038	1.7025	29.2162
82	0.0008	45.0268	2.7296	18.3539	132	0.0000	44.7026	1.6896	29.4388
83	0.0007	45.0084	2.6962	18.5739	133	0.0000	44.7014	1.6768	29.6614
84	0.0007	44.9909	2.6636	18.7941	134	0.0000	44.7002	1.6643	29.8840
85	0.0006	44.9744	2.6317	19.0143	135	0.0000	44.6992	1.6519	30.1067
86	0.0005	44.9587	2.6007	19.2347	136	0.0000	44.6981	1.6398	30.3293
87	0.0005	44.9439	2.5704	19.4551	137	0.0000	44.6972	1.6278	30.5520
88	0.0004	44.9299	2.5408	19.6757	138	0.0000	44.6963	1.6160	30.7747
89	0.0004	44.9166	2.5118	19.8963	139	0.0000	44.6954	1.6043	30.9974
90	0.0003	44.9040	2.4836	20.1171	140	0.0000	44.6946	1.5929	31.2201
91	0.0003	44.8921	2.4560	20.3379	141	0.0000	44.6938	1.5816	31.4429
92	0.0003	44.8808	2.4290	20.5588	142	0.0000	44.6931	1.5704	31.6656
93	0.0002	44.8701	2.4026	20.7798	143	0.0000	44.6924	1.5594	31.8883
94	0.0002	44.8600	2.3767	21.0009	144	0.0000	44.6917	1.5486	32.1111
95	0.0002	44.8504	2.3515	21.2220	145	0.0000	44.6911	1.5379	32.3339
96	0.0002	44.8414	2.3267	21.4432	146	0.0000	44.6905	1.5273	32.5567
97	0.0002	44.8328	2.3025	21.6645	147	0.0000	44.6899	1.5169	32.7794
98	0.0001	44.8247	2.2788	21.8859	148	0.0000	44.6894	1.5067	33.0022
99	0.0001	44.8170	2.2556	22.1073	149	0.0000	44.6889	1.4966	33.2250
100	0.0001	44.8097	2.2329	22.3288	150	0.0000	44.6884	1.4866	33.4478

Notes: Calculation of W/H has an error of less than 1%.

Source: Gupta, K.C., Garg, R., Bahl, I., Bhartia, P., *Microstrip Lines and Slotlines,* 2nd Ed., Artech House, Norwood, MA, 1996, 103.



$$L_{WIRE\ over\ GND} = \frac{\mu_0\mu_r}{2\pi}\cosh^{-1}\frac{h}{a}\ \text{(nH/m)} \qquad \text{for } h \gg a$$

FIGURE B.6 Inductance of a wire over a ground plane. When consistent units are used for the wire radius, a, and the height over the ground plane, h, the formula provides an estimate of the inductance per unit length in (nH/m).

Bondwires, Ribbons, Mesh

The assembly of semiconductor and hybrid integrated circuits for microwave and millimeter-wave frequencies generally requires the use of gold bondwires, ribbons, or mesh. The impedance of the interconnection must be accounted for in a good design. Unfortunately, there is no single accepted electrical model. The complexity required of the model will depend on the frequency of operation and the general impedance levels of the circuits being connected. At low frequencies and moderate to high impedances, the connection is frequently modeled as an inductor (sometimes in series with a resistor); at high frequencies, a full 3-D electromagnetic simulation may be required for accurate results. At intermediate points it may be modeled as a high impedance transmission line or as a lumped LC circuit. Note that the resistances of the rf interconnects should be included in the design of extremely low-noise circuits as they will affect the noise figure. In connecting a semiconductor die to package leads, it may also be necessary to model the mutual inductances and interlead capacitances in addition to the usual self inductances and shunt capacitance. Figure 13 illustrates one method of modeling bond wire inductance that has been shown adequate for many microwave applications. More sophisticated methods of modeling bond wires, ribbon or mesh are described in the references.

References

1. Grover, F.W., *Inductance Calculations*, Dover Publications, New York, NY Available through Instrument Society of America, Research Triangle Park, NC.
2. Caulton, M., Lumped Elements in Microwave Circuits, from Advances in Microwaves 1974, Academic Press, NY 1974, pp. 143–202.
3. Wadell, B. C., Transmission Line Design Handbook, Artech House, Boston, MA 1991, pp. 151–155.
4. Terman, F.E., *Radio Engineer's Handbook*, McGraw-Hill, 1943 pp. 48–51.
5. Caverly, R. H., "Characteristic Impedance of Integrated Circuit Bond Wires," IEEE Transactions on Microwave Theory and Techniques, Vol MTT-34, No. 9, September, 1986, pp. 982–984.
6. Kuester, E. F., Chang, D. C., "Propagating Modes Along a Thin Wire Located Above a Grounded Dielectric Slab," IEEE Transactions on Microwave Theory and Techniques, Vol MTT-25, No. 12, December, 1977, pp. 1065–1069.
7. Mondal, J. P. "Octagonal spiral inductor measurement and modelling for MMIC applications," Int. J. Electronics, 1990, Vol 68, No. 1, pp. 113–125.
8. MIL-STD-883E, "Test Method Standard, Microcircuits" US Department of Defense.
9. Gupta, K. C., Garg, E., Bahl, I., Bhartia, P., *Microstrip Lines and Slothlines*, 2nd Ed., Artech House, Norwood, MA, 1996, 103.

Index

D

E

F

N

O

P

Q

R

S

T

U

V

W

Z

Electrical Engineering

RF AND MICROWAVE APPLICATIONS AND SYSTEMS

Ubiquitous. That's a word commonly used when discussing cell phones, GPS navigation, and many other products that utilize RF and microwave technology. Items that used to be considered unusual have become a part of everyday life, so much so that most people can't imagine doing without them. Yet as essential as these items seem, soon they too will be obsolete and a whole new range of products requiring RF and microwave technology will be available. Helping you meet the needs of the future, **RF and Microwave Applications and Systems** covers a broad array of RF and microwave systems used for communication, radar, and heating applications. The book discusses the fundamentals, technologies, and applications at the heart of modern wireless communications. It contains 304 illustrations, 93 tables, and 372 equations in material contributed by leading experts from industry and academia.

In addition to new chapters on WiMAX 802.11g, fiber optic links, electronics for defense applications and ground penetrating radar, nearly every existing chapter features extensive updates and several were completely rewritten to reflect the significant changes in areas such as wireless PANs, satellite communication systems, RF and microwave avionics, system considerations, and underlying electromagnetic physics.

Features

- Provides an overview of commercial communications systems including avionic, medical, and military applications
- Explores past, current, and emerging cellular systems, navigation systems, and satellite-based systems and includes coverage of voice and data systems
- Examines military electronics, avionics, and military and automotive applications of radar
- Discusses systems considerations including thermal, mechanical, reliability, power management, and safety
- Elucidates engineering processes through coverage of corporate initiatives, cost modeling, and design reviews
- Delineates the underlying physics of electromagnetic propagation and interference
- Covers RF/microwave energy used for therapeutic medicine

7219

ISBN 0-8493-7219-4

90000

9 780849 372193

CRC Press
Taylor & Francis Group
an informa business
www.taylorandfrancisgroup.com

6000 Broken Sound Parkway, NW
Suite 300, Boca Raton, FL 33487
270 Madison Avenue
New York, NY 10016
2 Park Square, Milton Park
Abingdon, Oxon OX14 4RN, UK

www.crcpress.com