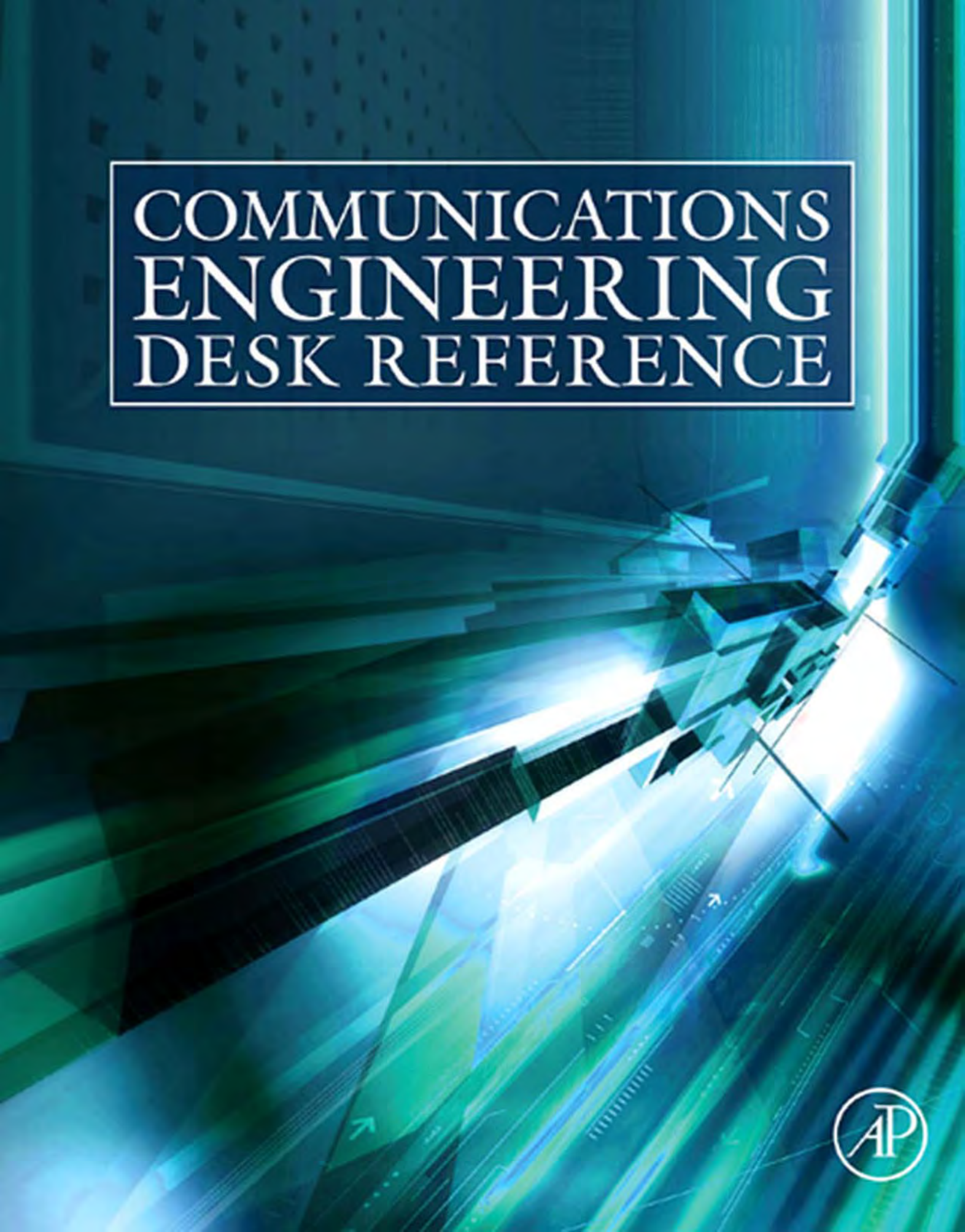
COMMUNICATIONS ENGINEERING DESK REFERENCE





Communications Engineering Desk Reference

Note from the Publisher

This book has been compiled using extracts from the following books within the range of Communications Engineering books in the Elsevier collection:

Dowla (2004) *Handbook of RF and Wireless Technology*, 9780750676953
Da Silva (2001) *High Frequency and Microwave Engineering*, 9780750650465
Fette (2006) *Cognitive Radio Technology*, 9780750679527
Kitchen (2001) *RF and Microwave Radiation Safety*, 9780750643559
Ellis, Pursell and Rahman (2003) *Voice, Video and Data Network Convergence*, 9780122365423
Van der Schaar and Chou (2007) *Multimedia over IP and Wireless Networks*, 9780120884803
Zhao and Guibas (2004) *Wireless Sensor Networks: An Information Processing Approach*, 9781558609143
Dahlman et al. (2007) *3G Evolution*, 9780123725332
Correia (2006) *Mobile Broadband Multimedia Networks: Techniques, Models and Tools for 4G*, 9780123694225
Dobkin (2005) *RF Engineering for Wireless Networks*, 9780750678735
Bensky (2004) *Short Range Wireless Communication*, 9780750677820
DeCusatis (2002) *Handbook of Fiber Optic Data Communication*, 9780122078910
Jack (2007) *Video Demystified*, 9780750683951
Bovik (2005) *Handbook of Image and Video Processing*, 9780121197926

The extracts have been taken directly from the above source books, with some small editorial changes. These changes have entailed the re-numbering of Sections and Figures. In view of the breadth of content and style of the source books, there is some overlap and repetition of material between chapters and significant differences in style, but these features have been left in order to retain the flavour and readability of the individual chapters.

End of chapter questions
Within the book, several chapters end with a set of questions; please note that these questions are for reference only. Solutions are not always provided for these questions.

Units of measure
Units are provided in either SI or IP units. A conversion table for these units is provided at the front of the book.

Upgrade to an Electronic Version
An electronic version of the Desk reference, the *Communications Engineering e-Mega Reference*, 9780123746498

- A fully searchable Mega Reference eBook, providing all the essential material needed by Communications Engineers on a day-to-day basis.
- Fundamentals, key techniques, engineering best practice and rules-of-thumb at one quick click of a button
- Over 1,500 pages of reference material, including over 1,000 pages not included in the print edition

Go to http://www.elsevierdirect.com/9780123746481 and click on **Ebook Available**

Communications Engineering Desk Reference



Amsterdam · Boston · Heidelberg · London · New York · Oxford
Paris · San Diego · San Francisco · Sydney · Tokyo
Academic Press is an imprint of Elsevier



Academic Press is an imprint of Elsevier
Linacre House, Jordan Hill, Oxford OX2 8DP, UK
525 B Street, Suite 1900, San Diego, CA 92101-4495, USA

First edition 2009

Copyright © 2009 Elsevier Inc. All rights reserved

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865 843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com. Alternatively visit the Science and Technology website at www.elsevierdirect.com/rights for further information

Notice
No responsibility is assumed by the publisher for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions or ideas contained in the material herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

ISBN: 978-0-12-374648-1

For information on all Academic Press publications visit our web site at elsevierdirect.com

Printed and bound in the United States of America

09 10 11 11 10 9 8 7 6 5 4 3 2 1

Working together to grow libraries in developing countries
www.elsevier.com | www.bookaid.org | www.sabre.org
ELSEVIER BOOK AID International Sabre Foundation

Contents

Author Biographies vii

Section 1 INTRODUCTION 1

1.0 Introduction 3

Section 2 RF ENGINEERING 13

2.1 Basic features of radio communication systems 15

2.2 Transmission lines 41

2.3 Software defined radio 69

2.4a The software defined radio as a platform for cognitive radio 91

2.4b Cognitive radio: The technologies required 115

2.5 Introduction to RF and microwave radiation 139

Section 3 NETWORK COMMUNICATIONS 151

3.1 Data and voice traffic 153

3.2 Network infrastructure 165

3.3 VoIP technology 189

3.4 Channel protection fundamentals 201

3.5 Network adaptive media transport 223

3.6 Real-time communication over networks 233

3.7 Wireless sensor networks 247

Section 4 MOBILE COMMUNICATIONS 257

4.1 Background of 3G evolution 259

4.2 The motives behind the 3G evolution 267

4.3 High data rates in mobile communications 273

4.4 OFDM transmission 281

4.5 Scheduling, link adaptation and hybrid ARQ 295

4.6 WCDMA evolution: HSPA and MBMS 305

4.7 Propagation modelling and channel characterisation 313

Section 5 SHORT RANGE WIRELESS COMMUNICATION 381

5.1 Wireless local area networks 383

5.2 Short-range wireless applications and technologies 413

Section 6 OPTICAL DATA COMMUNICATION 441

6.1 Optical fiber, cable and connectors 443

Section 7 VIDEO AND IMAGE PROCESSING 461
7.1 Introduction to video 463
7.2 Colour spaces 469
7.3 Video signals overview 483
7.4 Video compression 501

Section 8 APPENDIX 523
List of acronyms 525

Index 529

Author Biographies

Per Beming joined Ericsson in 1994 and he has worked with radio access concept development and standardization, primarily with architecture questions. The most prevalent examples are GPRS, WCDMA, HSPA, and LTE-SAE. He has been a key member of 3GPP TSG RAN since 1999. Currently Per is responsible for standardization at Developing Unit Radio within the Business Unit Networks, and is also manages a group of systems managers working with concept development and standardization.

Alan Bensky is an electronics engineering consultant with over 30 years of experience in analog and digital design, management, and marketing. Specializing in wireless circuits and systems, he has carried out projects for a variety of military and consumer applications and led the development of three patents on wireless distance measurement. He has authored two books on wireless communication and has written several articles in international and local publications. Bensky has taught electrical engineering courses, gives lectures on radio engineering topics and is a senior member of IEEE.

Dr. Alan Bovik is a Professor at the University of Austin in Texas in the Departments of Electrical and Computer Engineering, Computer Sciences and Biomedical Engineering. He is the inventor or co-inventor of, amongst other well-known inventions, Order Statistic Filters and the Image Modulation Model. He has published over 450 technical articles and holds two U.S. patents. Dr Bovik has received many awards, including the Technical Achievement Award and the Distinguished Lecturer Award of the IEEE Signal Processing Society. He is a Fellow of the IEEE and a Fellow of the Optical Society of America and has been involved in numerous professional society activities. He is also a registered Professional Engineer in the State of Texas and is a frequent consultant to legal, industrial and academic institutions.

Bruno Clerckx is currently Senior Engineer at Samsung Electronics in Korea. He has been working on 3GPP LTE RAN 1 and is currently involved in IEEE 802.16m. He previously held a visiting research position at the Smart Antennas Research Group (Information Systems Laboratory), Stanford University (CA, USA). He is the author of more than 30 research papers as well as multiple standard contributions. He also holds several Korean and international patents. He received the IEEE Symposium on Communications and Vehicular Technology Best Student Paper Award in 2002.

Dr. Philip A. Chou is a Principal Researcher with Microsoft Research in Redmond, Washington, and manages the Communication and Collaboration Systems research group. He has lectured at the Universities of Stanford, Washington, and the Chinese University of Hong Kong. He currently serves as coordinator of the awards subcommittee of the IEEE SP Society Multimedia Signal Processing technical committee (MMSP TC). He also serves on the editorial board of the IEEE Signal Processing Magazine. He is a Fellow of the IEEE and is the recipient of, amongst other awards, the Signal Processing Society Paper Award and the IEEE Transactions on Multimedia Best Paper Award.

Luis M. Correia is Professor in Telecommunications at Technical University of Lisbon. He has acted as a consultant for Portuguese GSM operators and the telecommunications regulator. He has authored many papers and communications in international journals and conferences, for which he has served also as a reviewer, editor, and board member. He has served as evaluator and auditor in ACTS, ESPRIT and IST frameworks, besides several national agencies worldwide, and was part of the COST Domain Committee on ICT. He was the Chairman of the Technical Program Committee of PIMRC'2002. He is part of the Expert Advisory Group and of the Steering Board of the European eMobility platform.

Dr. Erik Dahlman has been with Ericsson Research since 1993 and where he currently holds a position of Senior Expert in the area of Radio Access Technologies. He has been heavily involved in the development and standardization of 3G radio-access technologies, including WCDMA, HSPA, and LTE, both within local standardization bodies such as ETSI and ARIB, as well as within the global 3GPP organization.

Dr. Ed da Silva is a former Tenured Academic for the Open University Telematics Department and Professor and Head of the Electrical Department at Etisalat College. He spent over fifteen years in the industry in the USA, UK, Japan and Hong Kong, before spending over fifteen years in academia in the UK, United Arab Emirates and Nigeria. He also designed equipment for NASA's Apollo Moon Landings.

Dr. Casimer DeCusatis is an IBM Distinguished Engineer and Technical Executive based in Poughkeepsie, N.Y. He is an IBM Master Inventor with over 70 patents, and recipient of several industry awards, including the IEEE Kiyo Tomiyasu Award and the EDN Innovator of the Year. He is co-author of more than 100 technical papers, book chapters, and encyclopedia articles, and is also co-leader of the IBM Academy of Technology study "Innovation Ecosystems". He is a Fellow of the IEEE, Optical Society of America, and SPIE (the international optical engineering society), and various other professional organizations and honor societies.

Dr. Daniel Dobkin has been involved in the development, manufacturing, and marketing of communications devices, components, and systems for thirty years. He is the author of three books and about 30 technical publications, and holds 7 US patents as inventor or co-inventor. He has twenty years of experience in semiconductor process development, and ten years in radio engineering and RFID. He has taught classes on chemical vapor deposition, antenna principles, and radio-frequency identification in the US and Asia.

Professor Farid Dowla is a senior research engineer at the Lawrence Livermore National Laboratory (LLNL) and has been at LLNL since 1984. His present research interests include high-frequency radio-wave communications, and micro-wave and millimeter-wave radar imaging techniques. He also worked many years in the areas of seismic, acoustics, and biomedical signal processing. He has taught graduate and undergraduate electrical engineering courses in signal processing and communications at U.C. Berkeley and U. C. Davis.

Juanita Ellis has been at the forefront in working with corporations in the areas of Convergence, Computer Security and E-business. Some of these companies include Sony, JcPenney, SWBell, Boeing, Xerox, Bell Atlantic, MCI, Citibank and Toyota. Currently, she helps companies deploy Voice and Data networks, Converged Solutions, VPN Security and Call Center applications. She has also been a keynote speaker for Women in Technology and companies such as Cisco and SWBell. She has lectured for the University of Maryland European Division, Southern Methodist University, National Technology University and UCLA's Anderson School of Business and Engineering Extension Programs.

Dr. Bruce Fette is Chief Scientist in the Communications Networking Division business area of General Dynamics C4 Systems, working in LSI design, speech signal processing, advanced signal processing for telephony, and RF communications. He has 36 patents and has been awarded the "Distinguished Innovator Award". He has also worked with the Software Defined Radio (SDR) Forum from its inception, currently performing the role of Technical Chair, and is a panelist for the IEEE Conference on Acoustics Speech and Signal Processing Industrial Technology Track. Dr. Fette currently heads the General Dynamics Signal Processing Center of Excellence in the Communication Networks Division.

Professor Leonidas Guibas is Professor of Computer Science (and by courtesy, Electrical Engineering) at Stanford, having previously worked for Xerox PARC, Stanford, MIT, and DEC/SRC. At Stanford he heads the Geometric Computation group within the Graphics Laboratory and has developed new courses in algorithms and data structures, geometric modeling, geometric algorithms, sensor networks, and biocomputation. He is also part of the AI Laboratory the Bio-X Program, and the Institute for Computational and Mathematical Engineering. Professor Guibas is an ACM Fellow and winner of the ACM/IEEE Allen Newell award.

Keith Jack is Director of Product Marketing at Sigma Designs, a leading supplier of high-performance System-on-Chip (SoC) solutions for the IPTV, Blu-ray, and HDTV markets. Previously, he was Director of Product Marketing at Innovision, focused on solutions for digital televisions. Mr. Jack has also served as Strategic Marketing Manager at Harris Semiconductor and Brooktree Corporation. He has architected and introduced to market over 35 multimedia SoCs for the consumer markets.

Ronald Kitchen was with the Royal Air Force for fourteen years from the age of sixteen. He joined the Marconi Company in 1955 and was appointed Marconi Research Quality Manager and Radiation Protection Officer in 1971, carrying out just under 1000 radiation investigations. He retired in 1990 and set up his own consultancy. As a visiting consultant, he created a course at Marconi College, training 142 Military officers and Senior NCOs. He was awarded the British Empire Medal for services to GEC Marconi.

Dr. Wing-Kuen Ling has lectured at King's College London since 2004. He is the author of the textbook *Nonlinear Digital Filters: Analysis and Applications* and an editor of the research monograph C*ontrol Chaos for Circuits and Systems: A Practical Approach*. His research interests include symbolic dynamics, optimization theory and applications, filter banks and wavelets, and fuzzy and impulsive control theory and applications.

David P. Morgan is a Surface Acoustic Waves (SAW) consultant, with over forty years experience in this area. He has authored two books and has published over 100 technical papers. His knowledge of the SAW area has led to his being invited to lecture on the subject in the U.S., Russia, Finland, Japan, China and Korea. He is a Life Senior Member of the IEEE.

Dr. Claude Oestges is a Research Associate of the Belgian National Science Foundation (FRS) and a part-time Associate Professor at UCL. He co-authored *MIMO Wireless Communications* and has made more than 100 contributions to international journals and conference proceedings. He was a member of the IEEE 802.11 Standardization Working Group on *"Multiple antenna channel modeling"*. He received the IEE Marconi Premium Award in 2001 and the IEEE Vehicular Technology Society 2004 Neal Shepherd Award.

Ron Olexa has designed and developed cellular telecommunication systems in major US and European markets. He served as the COO of SCT, then moved into the emerging Wireless data industry as CTO of Advanced Radio Telecom. He then started an independent consulting company to provide RF and technical guidance to spectrum licensees and companies. More recently he implemented one of the world's first WiMax capable networks. He has served on the Board of Directors of EdgeFocus Inc and has authored training material used for certification by the WiMax forum.

Dr. Stefan Parkvall joined Ericsson Research in 1999 where he currently serves as senior specialist in adaptive radio access, working with research on and standardization of future cellular technologies including HSPA, LTE and LTE-Advanced. His previous positions include being an assistant professor in communication theory at the Royal Institute of Technology, Stockholm, Sweden, and a visiting researcher at University of California, San Diego, USA. Dr Parkvall is a senior member of IEEE and received Ericsson's *Inventor of the Year* award 2005.

Charles Pursell has worked in the telecommunications and data networking industry for over 20 years. He began his career at Bell Laboratories as a systems architect for PBX and data network connectivity solutions and has worked in many positions in voice, data, and converged network design and support. He is currently a converged networking consultant for Avaya, Inc.

Joy Rahman, an award-winning senior converged engineering specialist, started his career at Lucent Technologies, Bell Lab Innovations, supporting enterprise solutions that involved convergence, VoIP, wireless, gigabit switching, ATM, and Frame Relay LAN/WAN solutions. He currently designs and provides solutions on advanced and complex converged networks for Fortune 500 companies, educational institutions, and security firms.

Johan Skold has been with Ericsson Research since 1989, where he has been involved in the standardization and evolution of GSM, EDGE and UMTS/HSPA/LTE. Mr. Skold was active in the FRAMES project within the European 4th Framework program, and initiated the work on EDGE as a GSM evolution within that project and later in standardisation. He was also active in taking the WCDMA concept from the FRAMES project into the standardization of UMTS/IMT-2000 in 3GPP, and in the development of HSPA and LTE as the evolution of 3G. He is a Senior member of IEEE.

Professor Mihaela van der Schaar is Associate Professor at the UCLA Electrical Engineering Department. Before this she was a senior researcher at Philips Research in the Netherlands and USA. She has published extensively on multimedia communications, networking, architectures, systems, compression and processing, and holds 30 US patents. She is an active participant in the ISO Motion Picture Expert Group (MPEG) standard and chairs the ad-hoc group on MPEG-21 Scalable Video Coding. She is also a senior member of IEEE, and the Technical Committee on Multimedia Signal Processing. She has won several awards, including the Okawa Foundation Award and the IBM Faculty Award three times.

Dr. Feng Zhao is a Principal Researcher at Microsoft Research, where he manages the Networked Embedded Computing Group. He has taught at Stanford University and Ohio State University, was a Principal Scientist at Xerox PARC and directed PARC's sensor network research effort. He serves as the founding Editor-In-Chief of ACM Transactions on Sensor Networks, and has authored or co-authored over 100 technical papers and books. He has received a number of awards, and his work has been featured in news media such as BBC World News, BusinessWeek, and Technology Review.

This page intentionally left blank

Section **One**

Introduction

This page intentionally left blank

Chapter 1.0

1.0

Introduction

John T. Moring

The last two decades have been the most dynamic in the history of wireless communications [1]. Most notably, mobile voice communications has exploded from a tiny niche market to a part of our daily lives. Building on comparable advances in computers and networking technology, today's wide area and local area wireless systems are poised to take us to the next level of mobile data services, where all the capabilities of the Internet are literally at our fingertips wherever we go.

In this chapter, we briefly review the history of wireless communications, survey today's wireless landscape, and introduce some of the leading edge topics covered later in this volume [2].

A short history of wireless communication

Figure 1.0-1 shows a time line of the development of wireless communications. We are well into the second century of radio communications. The pioneering work of Faraday, Maxwell, Hertz, and others in the 1800s led to Marconi's wireless telegraph at the turn of the century. The precursors to mobile radio as we know it have been available since the first transportable voice radios of the 1920s. Radio technology matured in the subsequent decades, with broadcast radio and television, and the portable manpack walkie-talkies of World War II. In the 1940s, cellular technology was conceived, with the ability to divide radio frequency service areas into "cells" to reduce interference and increase capacity. This is the basis for today's wide area voice and wireless local area networking technologies. Within a few years of the first satellite launch in 1957, satellites were being sent into space to act as communication relays.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2004 Elsevier Inc. All rights reserved

In 1969, the Federal Communications Commission (FCC) allotted portions of the radio frequency spectrum for mobile telephones. In the 1970s the Post Office Code Standardization Advisory (POCSAG) numeric paging code was standardized, and AT&T rolled out the first mobile telephone services operating on a cellular system. In 1987, the FCC allowed the use of new technologies in the 800 MHz cellular spectrum, with the first digital cellular transmissions (code division multiple access [CDMA], time division multiple access [TDMA], and global system for mobile communication [GSM]) tested in the United States shortly thereafter. With the adoption of digital technologies, new features such as voice mail, fax, and short messages have been enabled.

The boom in wireless usage in the 1990s (paralleling the Internet boom) has led to near ubiquitous wireless voice service throughout the United States and in much of the world. Personal wireless data services, exemplified by such technologies as short message service (SMS), wireless application protocol (WAP), ReFlex, Bluetooth, i-Mode, and 802.11, offer a range of mobile data services that are not far behind. For every wireline technology, from serial cable to fiber optics, there is an analogous wireless technology available when it is not feasible or convenient to use a cable connection. Figure 1.0-2 depicts how rapidly newer technologies grew in the 1990s while the number of wireline telephone installations in homes remained relatively static.

Where we are

Today's wireless technologies offer an immense range of capabilities to the user. System throughputs continue to expand, offering the ability to support an increasing number of applications. Wireless communication system

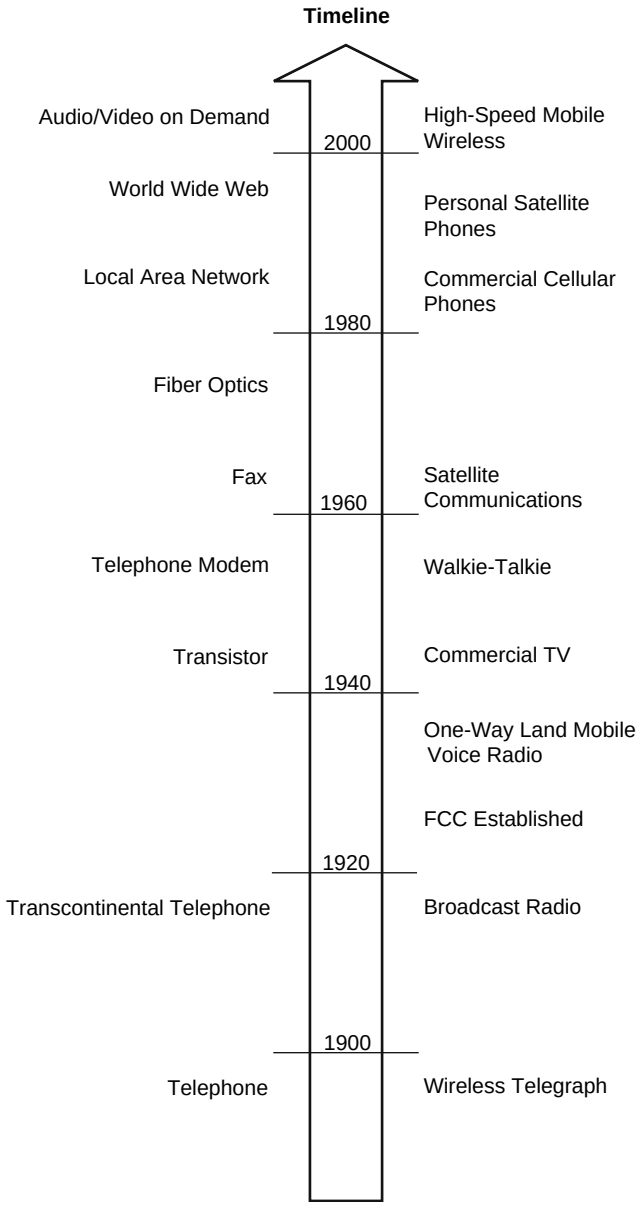


Figure 1.0-1 The graph indicates general telecommunications advances on the left and wireless-specific advances on the right.

availability is also increasing, due to investment in fixed infrastructure, as well as reduced device cost and size.

Figure 1.0-3 categorizes select wireless technologies, graphed by system throughput and user mobility. Several groupings are identified for convenience. On the left are *Fixed Location* systems, such as point-to-point microwave and stationary satellite systems, which generally operate at high rates (over one Mbps) on line-of-sight channels. Near the fixed systems, providing limited mobility over shorter transmission paths but still supporting Mbps data rates, are *Local Area* systems, such as wireless local area networks (802.11) and personal area networks (Bluetooth). Finally, *Wide Area Mobile* systems, such as paging and cellular, provide extended mobility but with relatively limited throughput. These categories are explored in the following section.

Before entering a discussion of specific wireless technologies, it is useful to review the relevant characteristics of a generic radio system. Figure 1.0-4 illustrates a wireless system, showing a signal sent from the transmitter on the left to the receiver in the center. Other aspects of the environment are shown to highlight the challenges inherent in wireless communications. These challenges are the focus of much research aimed at improving RF communications.

First, even in the best of situations, we have free space *attenuation*, where the signal loses strength at a rate proportional to the square of the distance traveled. This limits the signal propagation distance. The electromagnetic radio waves are further attenuated due to *blockage* by objects (including atmospheric particles) in their propagation paths. Both types of attenuation limit the ability of the receiver to capture and interpret the transmitted signal.

Additionally, radio signals are subject to reflection, which leads to *multipath* fading, another form of signal loss. A reflected signal, since it travels a longer distance between transmitter and receiver, arrives at the receiver with a time delay relative to a signal following a direct (or shorter reflected) path. Two similar signals, offset in time, may cancel each other out.

Another difficulty facing the receiver operating in the unprotected wireless environment is the possibility of other similar signals sharing the same medium and arriving at the receiver simultaneously with the desired signal, thus causing *interference*. Finally, the unprotected medium also allows the possibility of eavesdropping, or *interception*, where a third party captures the transmitted signal without permission, potentially compromising the privacy of the system's users.

Each of the challenges illustrated in Figure 1.0-4 identifies an area where wireless communications is at a disadvantage to a comparable wireline communication system. Why then are wireless communications so prevalent? Each wireless deployment may have its own design rationale, but two fundamental reasons cover most cases. First, wireless systems provide flexibility in deployment. Whether connecting a laptop PC in a conference room or a pipeline monitor in the Arctic, the setup of a radio connection may be orders of magnitude faster than a wireline connection, due to the fact that no physical connecting media is required. Second, wireless systems can provide the option of mobility. Communicating on the move is very useful, if not critical, to many applications, and this is just not well supported through a wired medium.

Note that some of these "weaknesses" can cleverly be turned to the user's advantage. For example, attenuation

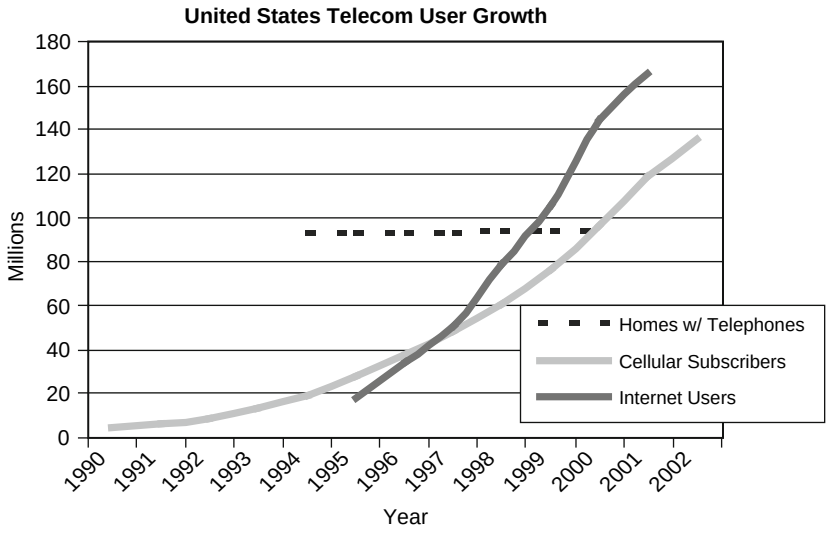


Figure 1.0-2 United States Telecom User Growth. With voice line penetration in saturation, wireless and Internet users continue to grow. Wireless data usage will follow. (Internet users include the United States and Canada.)

and blocking may be leveraged to limit signal propagation and opportunities for eavesdropping.

Having digressed far enough into the advantages and disadvantages of wireless, let us return to the discussion of how the wireless landscape is populated, starting with fixed location systems, and then addressing local area and wide area systems.

Fixed location systems

Fixed location systems by their nature can generally support high-gain stationary antennas and connections to the power grid; the resulting high signal-to-noise ratios provide the opportunity to operate at high throughput over long-range line-of-sight paths. Several classes of fixed system are described in the following, encompassing both RF and optical as well as terrestrial and space-based systems.

Point-to-point RF

Point-to-point microwave systems typically are used as a substitute for high-speed megabit (T-1, T-3) telecom circuits. They traditionally employ licensed frequencies above 2 GHz and user-owned end equipment. They use highly directional antennas that can span 10 miles and more, given line of sight. (To increase line of sight, systems are often located atop tall buildings and mountaintops.) A recent trend in this domain is toward

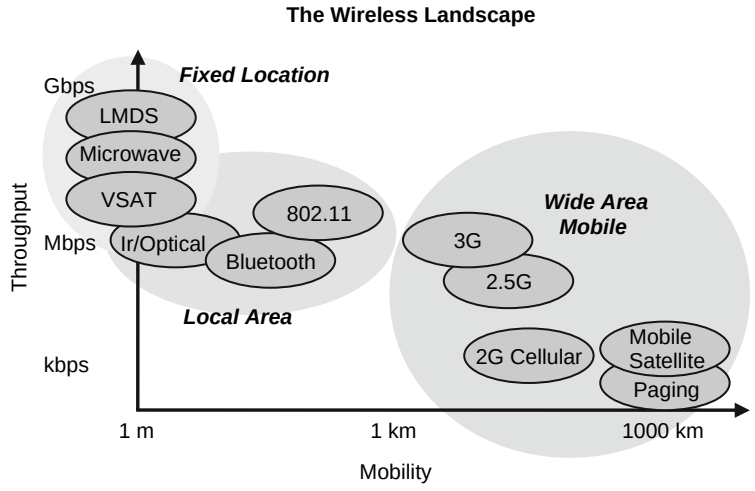


Figure 1.0-3 Current technologies in the wireless landscape provide a range of choices, from high-bandwidth fixed systems to wide area systems supporting low to moderate data rates.

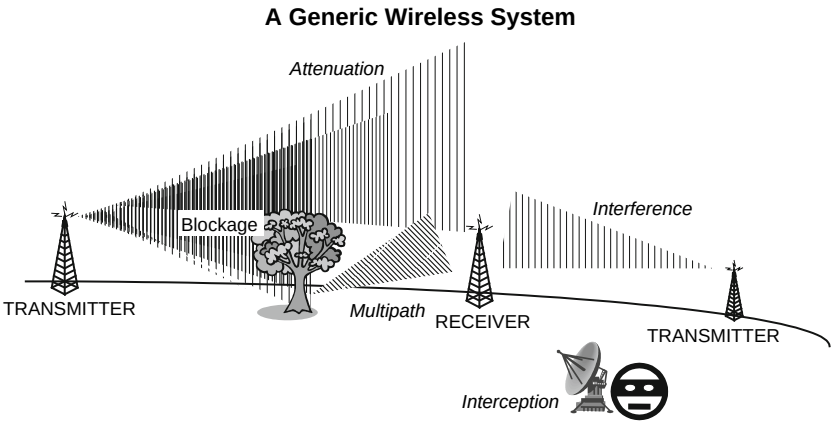


Figure 1.0-4 A generic wireless system. Inherent weaknesses of the wireless medium are offset by its flexibility and support for mobility.

lower-cost, unlicensed equipment, operating in the 2.4 GHz industrial, scientific, and medical (ISM) or 5 GHz unlicensed national information infrastructure (U-NII) bands.

Point to multipoint

Local multipoint distribution system (LMDS) and multichannel multipoint distribution service (MMDS) are carrier-grade technologies intended for wireless Internet access and general communication services. Spectrum has been allocated for these systems in several super high frequency (SHF) bands, offering tens of megabits per second of throughput. Along these lines, the Institute of Electrical and Electronic Engineers (IEEE) 802.16 working group is developing a family of standards for wireless metropolitan area networks (MANs). The technology provides a competitor to cable modem and DSL access, with additional flexibility. The current telecom slump has slowed the deployment of these systems.

VSAT and other geosynchronous satellite systems

Very small aperture terminal (VSAT) systems are similar to fixed terrestrial multipoint systems, except that instead of a tower- or rooftop-mounted base unit, a satellite transponder is used. A limited number of geosynchronous earth orbit (GEO) satellites circle the equator at 22,236 miles altitude, maintaining a fixed position in relation to a point on the earth. These transponders can be used for high-bandwidth bidirectional signal relay, supporting applications such as data transfer, Internet access, and audio/video signal distribution. Though the ground station antenna is fixed during operation, it is easily deployable, making it well suited for disaster recovery and other temporary situations.

Again, multiple licensed bands are employed. Full satellite channels are often subdivided via multiple access techniques and made available by resellers in the range of 20 kbps to 2 Mbps. According to the Global VSAT Forum [3], there are over 500,000 VSAT terminals installed worldwide.

Free space optical

A fairly recent arrival on the scene is free space optical (FSO) communications, which shares characteristics with both point-to-point RF and fiber optic technologies. As in point-to-point systems, a focused signal carries high-throughput bitstreams between two points; as in fiber optics, a laser or light-emitting diode (LED) is used to generate the optical signal that carries the information.

Operating with no spectrum license, and with speeds in excess of 1 Gbps, FSO offers an attractive choice for some backhaul and LAN extension applications. One weakness in today's systems is their susceptibility to optical blockage, particularly from fog.

Local area systems

Unlike the fixed systems just considered, local area systems achieve their high throughput via proximity. They generally allow some range of motion on the user's part, providing flexible usage scenarios. The technologies considered here require no spectrum license.

Infrared, IrDA

Infrared signals are used for a range of simple control protocols, such as TV remote controls. The Infrared Data Association (IrDA) has standardized multimegabit-per-second data transfers in support of a wide range of usage profiles. Typically, these are very short-range (i.e., inches to feet of separation between units) applications requiring a fairly high degree of directionality between transmitter and receiver. According to IrDA [4], over 300 million devices support this technology.

Bluetooth personal area networks

Operating at 1 Mbps channel rate in the unlicensed 2.4 GHz ISM band, Bluetooth (named for Harald Bluetooth, an ancient Danish Viking king) is intended for low cost and high interoperability. It is variously described as a personal area network (PAN) or a cable replacement technology, and should eventually be routinely embedded in cell phones, computers, PDAs, printers, and a plethora of other products that today require a cable for communications. Over time, more applications (as defined by Bluetooth "Profile" specifications) will be available in such diverse areas as video conferencing and automotive support. With many big names behind it (Bluetooth Promoter companies are 3Com, Agere, Ericsson, IBM, Intel, Microsoft, Motorola, Nokia, and Toshiba) and over 900 distinct Bluetooth [5] products (as of this writing), Bluetooth has a lot of momentum. It also has detractors and competitors.

Wireless local area networks

The big news in wireless communications over the last year or two has not been the much-hyped third generation (3G) cellular, but rather wireless local area networking (WLAN) technologies. These technologies (mostly based on the IEEE 802.11 family of standards and marketed under the Wi-Fi banner) operate in unlicensed spectrum, and provide wired LAN extension or replacement. The Wi-Fi family, also known as *Wireless Ethernet*, allows generic applications to operate seamlessly over a transparent layer 1 and 2.

Initially deployed as corporate LAN extensions, their low price, high performance (to tens of Mbps), and ease of operation have made them popular in the home office environment and to a lesser degree as a substitute for traditional point-to-point microwave systems. Additional vertical applications are starting to appear. Some vehicular toll collection systems employ 802.11, and commercial airlines have announced installation of WLANs in their aircraft for in-flight passenger use. Free access community networks are sprouting in numerous high-tech locales, potentially challenging the business model of established wireless carriers. Wi-Fi's well-publicized security limitations are expected to be solved before they significantly slow adoption of the technology. Over 26 million Wi-Fi devices are expected to ship in 2003 [6].

Wide area systems

Wide area mobile systems generally sacrifice throughput for longer propagation paths, user mobility, decreased device size, and increased coverage. Paging and cellular communication systems are prime examples of wide area systems.

Paging

Paging was successful for years as a reliable delivery mechanism for short text strings, such as phone numbers. With simplex data delivery, strategically placed high-power transmitters and robust encoding, excellent reliability was achieved. To compete with evolving cell phone services in the 1990s, many paging systems were upgraded to provide two-way transfers and larger message sizes, making them similar in function to the terrestrial packet systems described in the following. Now the Motorola proprietary two-way paging protocol ReFlex is the most widely deployed paging technology.

Faced with the increased coverage and penetration of cellular service, paging subscribership peaked in 1999 [7]. Motorola, the traditional leader among paging manufacturers, has announced that it is transitioning away from both one- and two-way paging equipment in favor of advanced cellular products and services [8].

Mobile satellite services

The 1990s saw the design of a number of LEO and MEO (low and medium earth orbit) satellite systems. Of these, only a few were actually launched, and these have had trouble maintaining financial viability in the face of ever-increasing terrestrial cellular coverage. Two systems provide cellular-like service: Globalstar and Iridium. Others (e.g., Orbcomm) provide two-way paging-like services (though with increased latency, as its LEO satellites are not always overhead). More ambitious systems such as Teledesic await the day when the investment climate may be more propitious.

Successful service providers leverage mobile satellite technology to benefit niche markets. Inmarsat uses a GEO satellite constellation to provide connectivity to maritime and other high-value assets. Omni-Tracs provides vehicular tracking and communications via satellite to the long-haul trucking industry.

Specialized private systems

Most wide area radio communications—like cellular and paging—are provided as services offered by carriers to the public. A notable exception is military systems, which typically use specialized proprietary designs in restricted frequency bands. Another exception is specialized private systems, used for such applications as voice dispatch and remote asset monitoring. These systems operate in licensed frequency bands and generally consist of several tower sites and many fixed or mobile subscriber units operated by a corporate or government entity. They go by many different names: SMR (specialized mobile radio), MAS (multiaddress system), trunked radio, PMR (private mobile radio), and others. Without the economies of scale of their mass market

counterparts, the specialized radio systems are often a generation behind comparable commercial equipment, with many still employing analog modulation, for example, long after cellular has migrated to digital.

Terrestrial packet systems

Several proprietary systems have been deployed to provide general-purpose packet data services. Mobitex, originally developed by Ericsson, and Ardis, now operated by Motient, provide shared channels operating at about 10 to 20 kbps. Wireless e-mail (similar to two-way paging) has been the dominant application for these systems.

The Ricochet system operates somewhat differently. With low power transceivers mounted on light poles located roughly every quarter mile, the system uses mostly unlicensed spectrum to offer 100-plus kbps Internet access to subscribers. This proprietary system was deployed in 15 or 20 markets before being turned off for financial reasons. At the time of this writing, it is back online in two markets.

Cellular

The flagship wireless service is unquestionably cellular. From a few million North American users 10 years ago, the number of mobile subscribers exceeded the number of fixed telephone subscribers by the end of the 1990s as shown in Figure 1.0-2. Analog service offered by two carriers per market has given way to highperformance, feature-rich digital services (including SMS, caller ID, three-way calling, etc.), often with four or more carriers per market; additional base stations have been installed to the point where dropped calls and dead spots are exceptions to the rule of ubiquitous connectivity.

North American consumers still have to choose between three or four incompatible second generation digital technologies, though the evolution path points toward two technologies in the coming generation (3G). Within the last year, cellular carriers have deployed their first serious large-scale data offerings, with medium rate (~50 kbps) packet data overlaid on their voice channels. For the first time, wide area consumer data services are available at moderate data rates and low costs. It is yet to be seen exactly how this offering will be used, or how widely it will be accepted.

Applications

We just discussed the current wireless landscape from a technological viewpoint. We can also consider wireless products in terms of the supported applications, such as voice, messaging, private data, and Internet access.

Voice

Cellular is the most obvious wireless voice technology. SMR and satellite systems support vertical voice markets, and various technologies (including Bluetooth) are used for short-range cordless telephony. Efforts are underway to support voice services over WLAN technologies.

Messaging

Communication via short wireless messages has echoed the popularity of e-mail in a wired environment. Most usage of the Mobitex, Ardis (packet services), and ReFlex (paging) systems today consists of messaging traffic. Terrestrial and satellite voice systems support the SMS, which carries text messages up to about 150 characters over the cellular network. Its successors, enhanced messaging service (EMS) and multimedia messaging service (MMS), now carry enhanced content such as pictures on some networks.

Private data

On an enterprise scale, VSAT and microwave technologies exemplify high-performance data transfer. On a personal scale, technologies such as Bluetooth provide a medium for private data transfer over short ranges. Virtual private network (VPN) is an example of software that allows private communications across a shared infrastructure (e.g., the Internet).

Internet access

Until recently, wireless Internet access involved a 10 kbps cellular channel and a limited web view using the text-based wireless application protocol (WAP) or similar mechanism. Today's digital cellular and wireless LAN services now allow users full web access at reasonable speeds over a wide area. For higher rate fixed access, VSAT services, and in some areas fixed multipoint systems, are available.

Where we are going

The advances of the recent decades show no sign of slowing. Despite the current (2003) telecom slump, wireless research and development continues apace. Incremental improvements in all facets of wireless communications should continue for the foreseeable future. Additionally, there are certain ongoing research areas that could potentially provide quantum advances across the wireless landscape.

Software radio

In conventional radios, we see open system interconnect (OSI) Layer 1 (physical) implemented in (mostly analog) hardware, with filters and amplifiers, for example,

designed for operation at specific frequencies consistent with the radio's intended usage. Layer 2 (data link) may be implemented in custom digital hardware or a general-purpose digital signal processor; higher layers are usually executed by software/firmware in general-purpose microprocessors. The block diagram of a conventional radio is shown in the top section of Figure 1.0-5.

The idea behind software radio is to move as much of the radio functionality (including the physical layer) as possible into software. However, performing the high rate sampling and signal processing required for a fully software radio has not yet proven to be commercially feasible. More recently, the related concept of software-defined radio has emerged as an R&D topic. In this case, dynamically reconfigurable logic arrays execute specific computationally intensive functions, such as coding or encryption, as needed. A logic block may be configured as a decoder during reception, then reconfigured as an encoder during the next transmit cycle. The lower section of Figure 1.0-5 shows a block diagram of a software-defined radio.

For both software radio or software-defined radio, the advantages are similar:

- *Flexibility*—the radio no longer needs to be designed with foreknowledge of the exact characteristics of the target usage.
- *Upgradability*—operation with a new radio technology can be achieved simply with a new software download.
- *Cost*—using today's technologies, digital logic gates are inherently less expensive than analog components.
- *Power consumption*—there is potentially more opportunity to optimize for lower power consumption (and therefore longer battery life) for a software /digital function than for analog.

Ultrawideband signals

Ultrawideband (UWB) has been promoted as a technology for applications ranging from high-speed communications to location determination to short-range ground-penetrating imaging. With pulse modulation and bandwidths in excess of 1 GHz, UWB potentially takes the advantages of spread spectrum one step farther. The intent is to operate UWB transmitters in an unlicensed mode overlapping licensed frequencies. After significant study, the FCC issued rules allowing limited UWB operation, but questions still remain in the minds of spectrum license holders who are concerned about interference.

Smart antennas

"Smart" antennas have been used for years in specialty applications, especially by the military. There are continuing efforts to develop commercially attractive solutions for applications such as cellular base stations and wireless LAN hot spots. Smart, or phased array, antennas

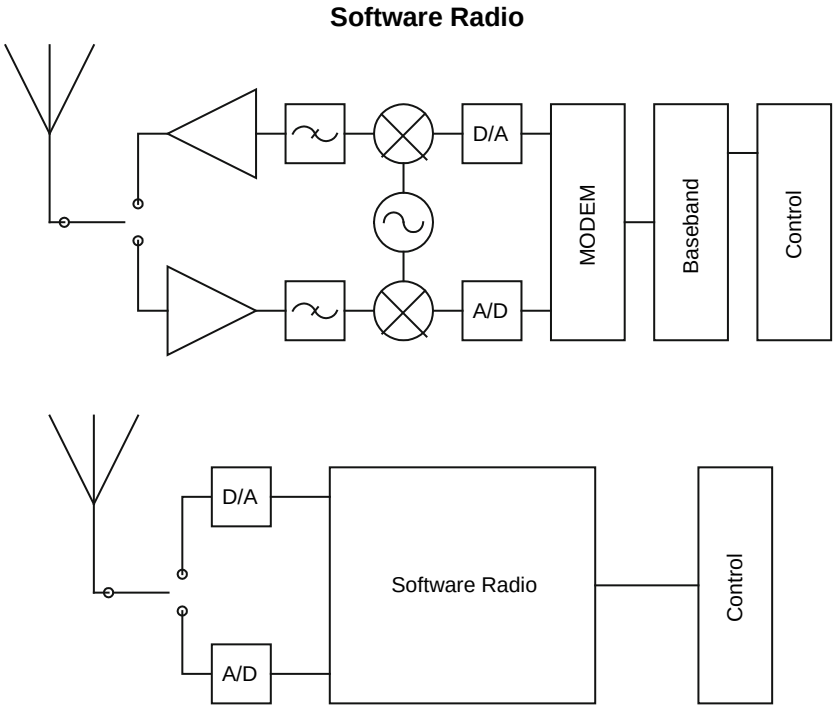


Figure 1.0-5 The ideal software radio (lower diagram) is much simpler and more flexible than the traditional radio (upper diagram), with its hardware tightly coupled to a specific signal type.

may be considered an extreme form of antenna diversity. An antenna array is composed of a collection of individual antenna elements, each of which is capable of signal reception (or transmission). A signal arriving at the array will typically be picked up at each element, offset slightly in time. By internally adjusting the phase (time offset) of each received signal and then summing, the antenna pattern is effectively focused on the remote transmitter. Signals arriving from other directions do not benefit from the summing and sink into the noise. The same technique may be used to benefit the transmitted signal as well as the received signal. One of the challenges of the technology is to dynamically adjust the offsets to maintain focus on a remote mobile unit. Figure 1.0-6 shows a simplified model of smart antenna operation.

Advanced networking

There are a number of advances at OSI layers 2 and above that will benefit wireless users, some examples of which are touched upon here. Over the past 15 years, deployment of the Internet protocol (IP) has advanced to where it is now a matter of course to connect from any computer to almost any other computer around the world, and to do it easily and quickly. This expectation also holds true for telephone-to-telephone connectivity. In some respects, mobile telephony is more advanced than IP networking in that a user can receive a call to his or her unique identifier (telephone number, including the home area code) while roaming to a different market with a different area code. This is not true today in data networking, in which a computer's unique identifier (IP address) is tied to a specific "location" on the network (i.e., a specific router). The situation is addressed by assigning a new IP address as the user reattaches through a different router. This presents limitations to some applications, constraining their ability to maintain a connection as their supporting device moves.

A set of features collectively known as *mobile IP* offers a solution to this problem, providing services to the mobile data use that is comparable to those provided by the mobile telephony infrastructure.

As mentioned earlier, security is one potential weakness of wireless systems. First generation cellular networks were vulnerable to theft of services as well as eavesdropping. These oversights were corrected with authentication and encryption systems built into the second generation systems. Likewise, standards will shortly address the security weaknesses of original IEEE 802.11 wireless LANs.

Another major area of research is in the field of quality of service (QoS). Applications (e.g., voice or video) that work well over dedicated wired links often experience difficulties when faced with the variable delays, channel errors, and throughput inherent in many wireless links. Wireless QoS features will provide a more consistent platform on which to support these applications.

In addition to specific research areas exemplified by those just listed, there are several more general trends that will affect the wireless landscape over the coming years.

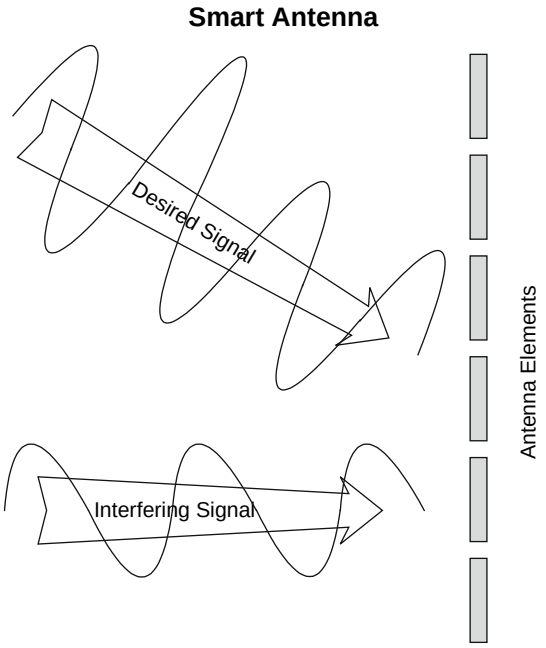


Figure 1.0-6 Smart antenna. In this simple example, the desired and interfering signals are captured at each antenna element. When aligning the received signals to maximize the desired signal, the offset versions of the interferer tend to cancel each other out.

Increased throughput

Apart from some low-speed niche applications such as telemetry, each wireless technology is continually pushing for higher data rates. Sophisticated modulation and coding can squeeze more data onto existing channels. Additionally, there is a continuing migration to higher frequencies, where there is more bandwidth and thus the potential for higher throughput. Compare for example the newer 802.11a, which supports 54 Mbps on each of up to 12 simultaneous channels at 5 GHz, with 802.11b, which offers 11 Mbps on each of three simultaneous channels at 2.4 GHz (under U.S. regulations). Recent news items report that additional spectrum is being proposed to supplement what is already designated for unlicensed use in the 5 GHz band.

Increased access

We have grown to expect near-ubiquitous coverage from our cellular phone carriers. This has been achieved through network tuning and intensive investment in network infrastructure. Next generation technology and

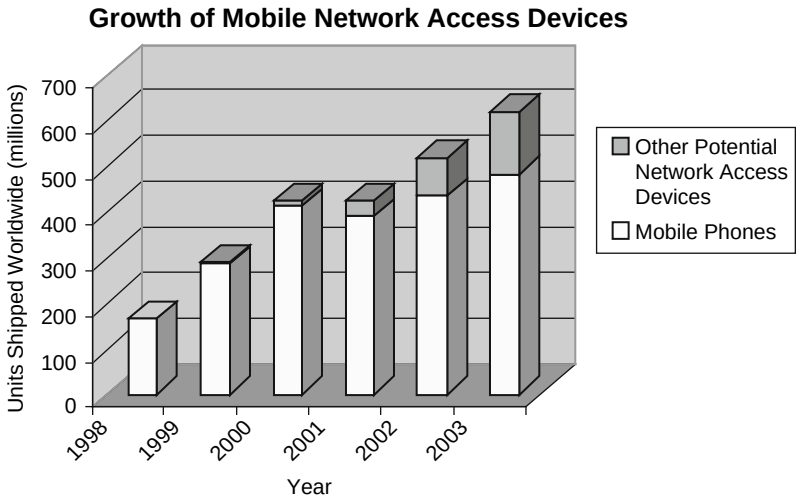


Figure 1.0-7 Growth of mobile network access devices [9]. "Other" devices include PDAs, handheld and laptop PCs, etc.

business consolidations should provide even more international roaming capabilities. Where there are several noninteroperable cellular technologies today, there are only two contenders for widespread international deployment in the third generation, and cellular operators are ensuring a much greater degree of compatibility between them. Additionally, the burgeoning growth of public and private Wi-Fi hot spots offers mobile data users another connectivity option. Bluetooth users carry their own PAN around with them wherever they go. And it is expected that satellite services will provide increasing connectivity options for those hard-to-reach remote locales.

Ubiquity

Not only is network coverage increasing, but the number of wireless devices is growing at a high rate. The reduced size and price of radios makes it feasible to add wireless capability to almost any device. Advances in supporting technologies, as well as wireless technologies, make radios viable where they never have been before.

- Battery power density continues to improve, reducing the size and increasing the utility of portable devices. Viable alternate power sources are on the horizon.
- User-interface advances (e.g., voice recognition, new display technologies) add convenience and potentially open the door for new applications.
- Integrated circuits continue to double their capabilities every year or two, allowing designers to pack more functions into ever-smaller packages. Newer semiconductor technologies and manufacturing processes provide increased efficiency.
- Processing techniques, such as software-defined radio, scripting and presentation languages (Java 2 Platform, Micro Edition [J2ME], binary runtime environment for wireless [BREW], extensible markup language [XML]), and video compression algorithms, also have the end result of providing more capability—and more configurable capabilities—to our mobile devices.

We are now seeing not only wireless phones and laptop computers, but also radio-connected PDAs, cameras, watches, and cars. Any device that has an embedded microprocessor is a candidate for wireless connectivity. Figure 1.0-7 illustrates the growth in mobile network access devices.

More applications

The increased availability of network coverage and wireless devices, and advances in associated electronics technologies, makes new applications viable. Mobile voice and mobile text messaging have been wildly popular. The marriage of cameras and wireless opens the door for some form of photo messaging. Cellular carriers (at the behest of the FCC, for safety reasons) are deploying location determination capabilities, some using the global positioning system (GPS). Once these capabilities are in place, a range of new applications beyond public safety will be available. Many see electronic gaming as a huge growth area, just now in its infancy.

Conclusion

The rich history of progress in wireless and RF communication has given us an array of technologies suited to a wide range of applications. We are in a time of explosive growth in wireless technology, as evidenced by the advances described throughout this book. Coupled with the

Internet content that is now available, the door is open for new applications ranging from telemetry to video. The 1990s saw a proliferation of wireless voice services; now, the wireless data boom will be even more dynamic because the types of devices and content are much more varied.

References

1. Coincidently, this period corresponds to the time this author has been employed as a communication engineer. Unfortunately, I can take only partial credit for the amazing advances of this era!
2. Many thanks to the University of Wisconsin in Madison and Melange Solutions in San Diego, for whom some of this material was originally developed.
3. http://www.gvf.org.
4. http://www.irda.org.
5. http://qualweb.bluetooth.org.
6. Dataquest, quoted by Krazit, Tom in "Microsoft Joins Home Wi-Fi Announcements," PCWorld.com (http://www.pcworld.com/news/article/0,aid, 10517,00.asp), September 19, 2002.
7. FCC, *Annual Report and Analysis of Competitive Market Conditions with Respect to Commercial Mobile Services: Sixth Report, July 17, 2001.*
8. Motorola, "Motorola Personal Communications Sector (PCS) Refocuses Messaging Division" (press release), December 3, 2001.
9. Adapted from the Shosteck Group white paper, *Lessons from Metricom and MobileStar: Success Factors for the Portable Internet Access Market* (January 2002).

Section **Two**

RF engineering

This page intentionally left blank

Chapter 2.1

2.1

Basic features of radio communication systems

Ed da Silva

2.1.1 Introduction

This chapter describes communication systems which use radio waves and signals. Radio signals are useful for two main reasons. They provide a relatively cheap way of communicating over vast distances and they are extremely useful for mobile communications where the use of cables is impractical.

Radio signals are generally considered to be electromagnetic signals which are broadcast or radiated through space. They vary in frequency from several kilohertz[1] to well over 100 GHz (10^{11} Hz). They include some well known public broadcasting bands: long-wave (155–280 kHz), medium-wave (522–1622 kHz), short-wave (3–30 MHz), very high frequency (VHF) FM band (88–108 MHz), ultrahigh frequency (UHF) television band (470–890 MHz) and the satellite television band (11.6 to 12.4 GHz). The frequencies[2] quoted above are approximate figures and are only provided to give an indication of some of the frequency bands used in Radio and TV broadcasting.

2.1.1.1 Aims

The aims of this chapter are to introduce you to some basic radio communications principles and methods. These include **modulation** (impressing signal information on to radio carrier waves), **propagation** (transmission of radio cartier waves) and **demodulation** (detection of radio carrier waves) to recover the original signal information.

The method we use here is to start with an overview of a communication system. The system is then divided to show its sub-systems and the sub-systems are then expanded to show individual circuits and items.

2.1.1.2 Objectives

The general objectives of this chapter are:

- to help you understand why certain methods and techniques are used for radio frequency (r.f.) and high frequency communication circuits;
- to appreciate the need for modulation;
- to understand the basic principles of modulation and demodulation;
- to understand the basic principles of signal propagation using antennas;
- to introduce radio receivers;
- to introduce you to the requirements of selectivity and bandwidth in radio communication circuits.

2.1.2 Radio communication systems

2.1.2.1 Stages in communication

Let's commence with a simple communications example and analyse the important stages necessary for communication. This is shown diagramatically in Figure 2.1-1.

[1] One hertz (Hz) means 1 cyclic vibration per second: 1 kHz = 1000 cyclic vibrations per second, 1 MHz = 1 000 000 cyclic vibrations per second, and 1 GHz = 1 000 000 000 cyclic vibrations per second. The word Hertz is named after Heinrich Hertz, one of the early pioneers of physics.
[2] The frequencies quoted are for Europe. Other countries do not necessarily follow the exact same frequencies but they do have similar frequency bands.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2001, Ed da Silva. Published by Elsevier Ltd 2001. All rights reserved

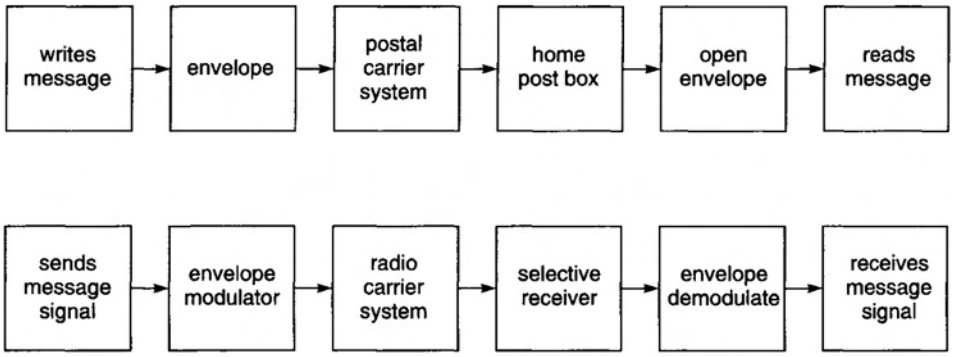


Fig. 2.1-1 Analogy between the postal system and a radio system.

We start by writing a *letter-message*, putting it in an *envelope*, and sending it through a *post-carrier* (postal carrier system) to our destination. At the other end, our recipient *receives* the letter from the post office, *opens* the envelope and *reads* our message. Why do we carry out these actions?

We write a letter because it contains the information we want to send to our recipient. In radio communications, we do the same thing; we use a message signal, which is an electrical signal derived from analog sound or digitally encoded sound and/or video/data signals, as the information we want to convey. The process of putting this information into an 'envelope' for transmission through the carrier is called modulation and circuits designed for this purpose are known as modulation circuits or **modulators**.

We use the post office as the carrier for our letters because the post office has the ability to transmit messages over long distances. In radio communications, we use a **r.f. carrier** because a radio carrier has the ability to carry messages over long distances. An r.f. carrier with an enveloped message impressed on it is often called an **enveloped carrier wave** or a **modulated carrier wave**.

When the post office delivers a letter to a destination, the envelope must be opened to enable the message to be read. In radio communications when the enveloped carrier wave arrives at its destination, the enveloped carrier must be 'opened' or **demodulated** to recover the original message from the carrier. Circuits which perform this function are known as demodulation circuits or **demodulators**.

The post office uses a system of postal codes and addresses to ensure that a letter is selected and delivered to the correct address. In radio communications, **selective** or **tuned circuits** are used to select the correct messages for a particular receiver. Amplifiers are also used to ensure that the signals sent and received have sufficient amplitudes to operate the message reading devices such as a loudspeaker and/or a video screen.

In addition to the main functions mentioned above, we need a post box to send our letter. The electrical equivalent of this is the **transmitting antenna**. We require a letter box at home to receive letters. The electrical equivalent of this is the **receiving antenna**.

2.1.2.2 Summary of radio communications systems

A pictorial summary of the above actions is shown in Figure 2.1-1. There are three main functions in a radio communications system. These are: modulation, transmission and demodulation. There are also supplementary functions in a radio communications system. These include transmitting antennas,[3] receiving antennas, selective circuits, and amplifiers. We will now describe these methods in the same order but with more detail.

2.1.3 Modulation and demodulation

Before discussing modulation and demodulation, it is necessary to clarify two points: the modulation information and the modulation method.

In the case of a letter in the postal system, we are free to write our messages (modulation information) in any language, such as English, German, French, pictures, data, etc. However, our recipient must be able to read the language we use. For example it is useless to write our message in Japanese if our recipient can only read German. Hence the modulation information system we use at the transmitter must be **compatible** with the demodulation information system at the receiver.

Secondly, the method of putting information (modulation method) on the letter is important. For example,

[3] Antennas are also known as aerials.

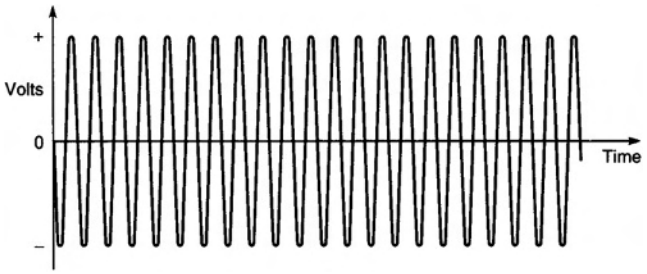


Fig. 2.1-2 A sinusoidal radio carrier wave.

we can type, use a pencil, ultraviolet ink, etc. However, the reader must be able to decipher (demodulate) the information provided. For example, if we use ultraviolet ink, the reader must also use ultraviolet light to decipher (demodulate) the message. Hence the modulation and demodulation methods must also be compatible.

In the discussions that follow we are only discussing modulation and demodulation methods; not the modulation information. We also tend to use sinusoidal waves for our explanation. This is because a great mathematician, Joseph Fourier,[4] has shown that periodic waveforms of any shape consist of one or more d.c. levels, sine waves and cosine waves. This is similar to the case in the English language, where we have thousands of words but, when analysed, all come from the 26 letters of the alphabet. Hence, the sinusoidal wave is a useful tool for understanding modulation methods.

We now return to our simple radio carrier wave which is the sinusoidal wave[5] shown in Figure 2.1-2.

A sinusoidal wave can be described by the expression

$$v_c = V_c \cos(\omega_c t + \phi_c) \qquad (2.1.1)$$

where

v_c = instantaneous carrier amplitude (volts)
V_c = carrier amplitude (peak volts)
ω_c = angular frequency in radians and $\omega_c = 2\pi f_c$ where
f_c = cartier frequency (hertz)
ϕ_c = carrier phase delay (radians)

If you look at Figure 2.1-2, you can see that a sinusoidal wave on its own provides little information other than its presence or its absence. So we must find some method of modulating our information on to the radio carrier wave. We can change:

- its *amplitude* (V_c) according to our information – this is called **amplitude modulation** and will be described in Section 2.1.3.1;

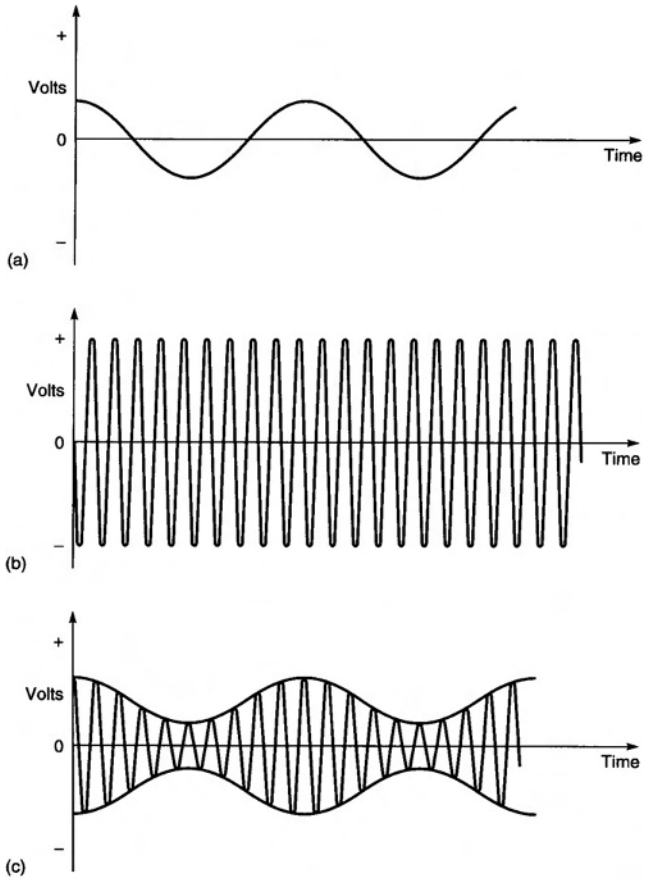


Fig. 2.1-3 Amplitude modulation waveforms: (a) modulating wave; (b) carrier wave; (c) modulated wave.

- its *frequency* (ω_c) according to our information – this is called **frequency modulation** and will be described in Section 2.1.3.2;
- its *phase* (ϕ_c) according to our information – this is known as **phase modulation** and will be described in Section 2.1.3.3;
- or we can use a combination of one or more of the methods described above – this method is favoured by **digital modulation**.

2.1.3.1 Amplitude modulation

Amplitude modulation (AM) is the method used in medium-wave and short-wave radio broadcasting. Figure 2.1-3 shows what happens when we apply amplitude modulation to a sinusoidal carrier wave.

Figure 2.1-3(a) shows the modulating wave on its own.[6] Figure 2.1-3(b) shows the carrier wave on its own.

[4] Fourier analysis will be explained fully in a later section.
[5] A sinusoidal wave is a generic name for a sine or cosine wave. In many cases, cosine waves are used because of ease in mathematical manipulation.
[6] I have used a cosine wave here because you will see later when we use Fourier analysis that waveforms, no matter how complicated, can be resolved into a series of d.c., sine and cosine terms and their harmonics.

Figure 2.1-3(c) shows the resultant wave. The resultant wave shape is due to the fact that at times the modulating wave and the carrier wave are adding (in phase) and at other times, the two waves are opposing each other (out of phase).

Amplitude modulation can also be easily analysed mathematically. Let the sinusoidal modulating wave be described as

$$v_m = V_m \cos(\omega_m t) \quad (2.1.2)$$

where

v_m = instantaneous modulating amplitude (volts)
V_m = modulating amplitude (peak volts)
ω_m = angular frequency in radians and $\omega_m = 2\pi f_m$

where

f_m = modulating frequency (hertz)

When the amplitude of the carrier is made to vary about V_c by the message signal v_m, the modulated signal amplitude becomes

$$[V_c + V_m \cos(\omega_m t)] \quad (2.1.3)$$

The resulting envelope AM signal is then described by substituting Equation 2.1.3 into Equation 2.1.1 which yields

$$[V_c + V_m \cos(\omega_m t)] \cos(\omega_c t + \phi_c) \quad (2.1.4)$$

It can be shown that when this equation is expanded, there are three frequencies, namely $(f_c - f_m)$, f_c and $(f_c + f_m)$. Frequencies $(f_c - f_m)$ and $(f_c + f_m)$ are called sideband frequencies. These are shown pictorially in Figure 2.1-4.

The modulating information is contained in one of the sideband frequencies which must be present to extract the original message. The bandwidth (bw) is defined as the highest frequency minus the lowest frequency. In this case, it is $(f_c + f_m) - (f_c - f_m) = 2f_m$ where f_m is the highest modulation frequency. Hence, a radio receiver must be able to accommodate the bandwidth of a signal.[7]

2.1.3.2 Frequency modulation

Frequency modulation (FM) is the modulation method used in VHF radio broadcasting. Figure 2.1-5 shows what happens when we apply frequency modulation to a sinusoidal carrier wave. Figure 2.1-5(a) shows the modulating wave on its own. Figure 2.1-5(b) shows the carrier wave on its own. Figure 2.1-5(c) shows the resultant wave. The resultant wave shape is due to the fact that the carrier wave frequency increases when the modulating signal is positive and decreases when the modulating signal is negative. Note that in pure FM, the amplitude of the carrier wave is not altered.

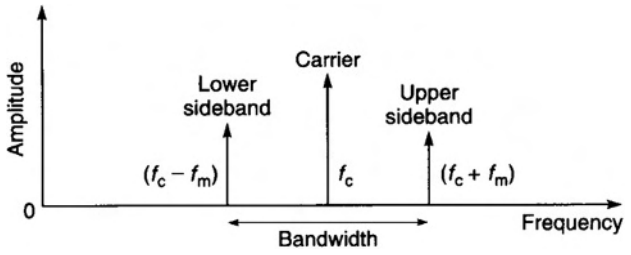


Fig. 2.1-4 Frequency spectrum of an AM wave.

The frequency deviation (Δf_c) of the carrier is defined as $[f_{c\,(max)} - f_{c\,(min)}]$ or

$$\Delta f_c = f_{c\,(max)} - f_{c\,(min)} \quad (2.1.5)$$

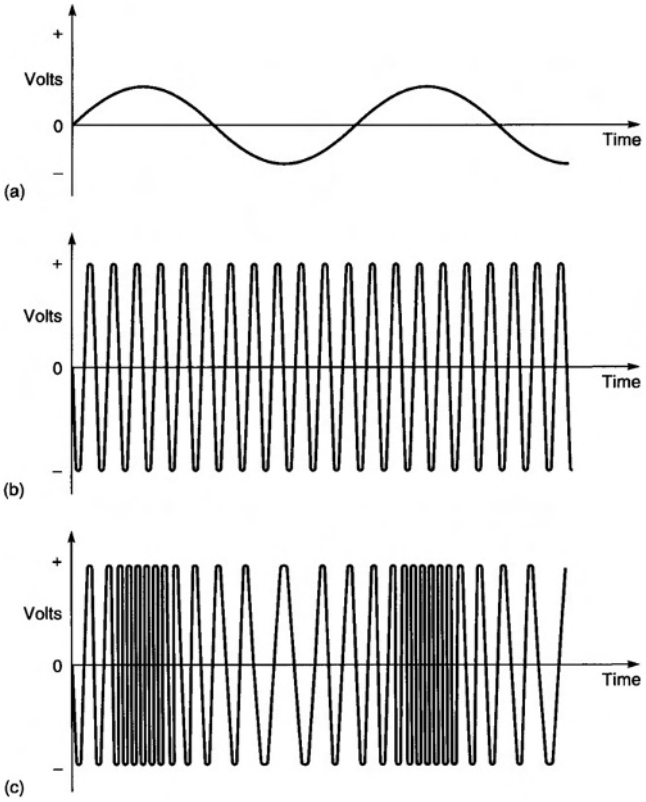


Fig. 2.1-5 Frequency modulation waveforms: (a) modulating wave; (b) carrier wave; (c) FM wave.

[7] This is not unusual because speech or music also have low notes and high notes and to hear them our own ears (receivers) must be able to accommodate their bandwidth. Older people tend to lose this bandwidth and often are unable to hear the high notes.

According to Carson's rule, the frequency bandwidth required for wideband FM is approximately 2 × (maximum frequency deviation + highest frequency present in the message signal) or

$$\text{bw} = 2[\Delta f_c + f_{m\,(max)}] \tag{2.1.6}$$

In FM radio broadcasting, the allocated channel bandwidth is about 200 kHz.

2.1.3.3 Phase modulation

Phase modulation (PM) is particularly useful for digital waveforms. Figure 2.1-6 shows what happens when we apply phase modulation to a sinusoidal cartier wave. Figure 2.1-6(a) shows a digital modulating wave on its own. We have used a pulse waveform as opposed to a sine wave in this instance because it demonstrates phase modulation more clearly. Figure 2.1-6(b) shows the carrier wave on its own. Figure 2.1-6(c) shows the resultant wave. Note particularly how the phase of the carrier waveform changes when a positive modulating voltage is applied. In this particular case, we have shown you a phase change of 180°, but smaller phase changes are also possible.

Phase modulation is popularly used for digital signals. Phase modulation is synonymous with frequency modulation in many ways because an instantaneous change in phase[8] is also an instantaneous change in frequency and vice versa. Hence, much of what is said about FM also applies to PM.

2.1.3.4 Combined modulation methods

Digital signals are often modulated on to a radio carrier using both phase and amplitude modulation. For example, an eight level coded digital signal can be modulated on to a carrier by using distinct 90° phase changes and two amplitude levels. This is shown diagrammatically in Figure 2.1-7 where eight different signals, points A to H, are encoded on to a radio carrier.

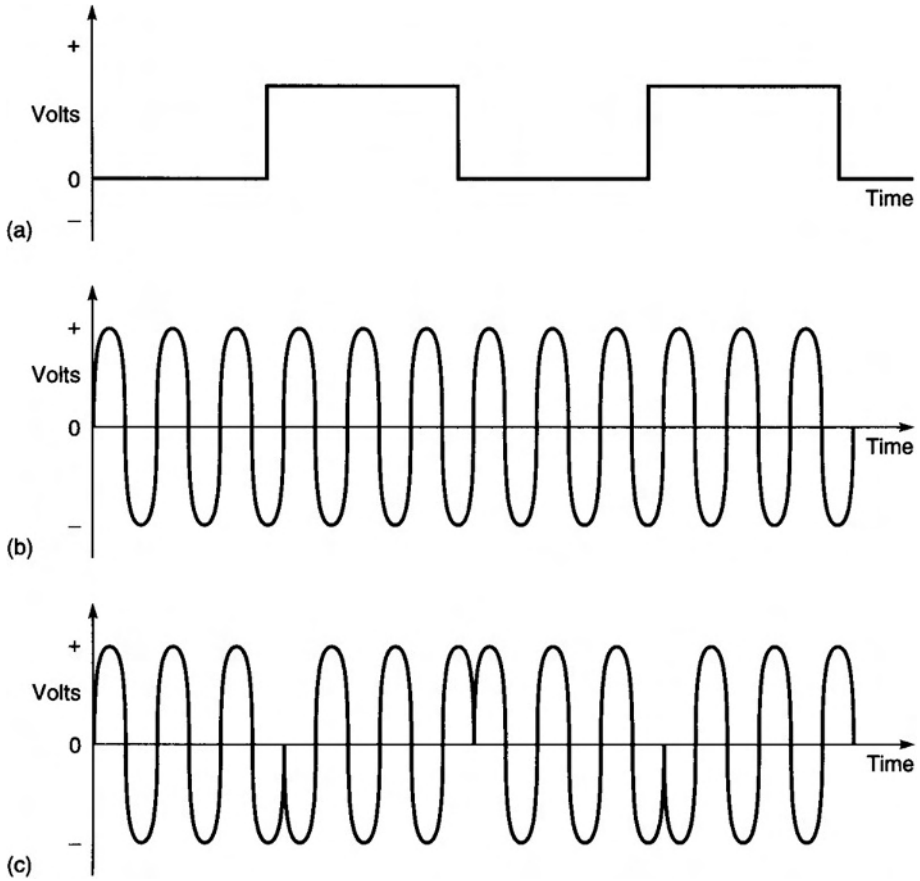


Fig. 2.1-6 Phase modulation waveforms: (a) modulating wave; (b) carrier wave; (c) modulated wave.

[8] Phase (ϕ) = angular velocity (ω) multiplied by time (t). Hence $\phi = \omega t$. Note this equation is similar to that of distance = velocity × time. This is because ϕ = amount of angle travelled = velocity (ω) × time (t).

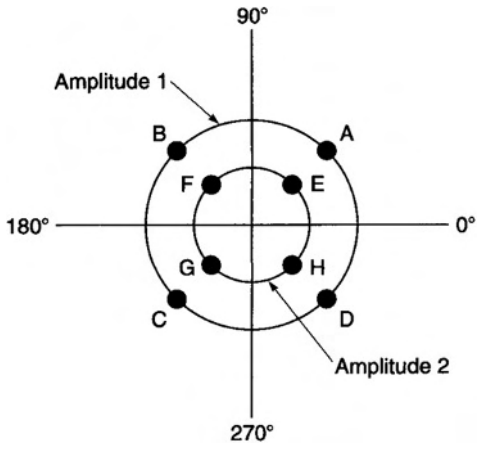


Fig. 2.1-7 An eight level coded signal modulated on to a radio carrier.

This method is also known as quadrature amplitude modulation (QAM).

2.1.3.5 Summary of modulation systems

In this section, we have shown you four methods by which information signals can be modulated on to a radio carrier.

2.1.4 Radio wave propagation techniques

2.1.4.1 Properties of electromagnetic waves

In Figure 2.1-8 we show the case of a radio generator feeding energy into a load via a two-wire transmission line. The radio generator causes voltage and current waves to flow towards the load. A voltage wave produces a voltage or electric field. A current wave produces a current or magnetic field. Taken together these two fields produce an **electromagnetic field** which at any instant varies in intensity along the length of the line.

The electromagnetic field pattern is, however, far from stationary. Like the voltage on the line, it propagates from end to end with finite velocity which – for an air spaced line – is close to the velocity of light in free space.[9] The flow of power from source to load is then regarded as that of an **electromagnetic wave** propagating between the conductors.

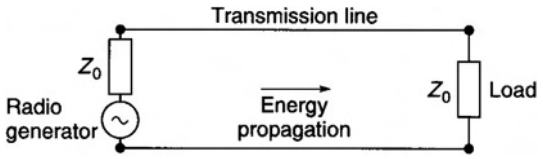


Fig. 2.1-8 Energy propagation in a transmission line.

The equivalence between the circuit and field descriptions of waves on transmission lines is demonstrated by the fact that at any point in the electromagnetic field the instantaneous values of the electric field (E) (volts/metre) and the magnetic field (H) (amperes/metre) are related by

$$\frac{E\ (\mathrm{V/m})}{H\ (\mathrm{A/m})} = Z_0\ (\mathrm{ohms}) \qquad (2.1.7)$$

where Z_0 is the characteristic impedance of the transmission line.[10] It can also be shown that both approaches give identical results for the power flow along a matched line.

In the two-wire transmission line shown in Figure 2.1-8, the parallel conductors produce electromagnetic fields which overlap and cancel in the space beyond the conductors. The r.f. energy is thus confined and guided by the conductors from the source to its destination. If, however, the conductor spacing is increased so that it becomes comparable with the wavelength of operation the line will begin to radiate r.f. energy to its surroundings. The energy is lost in the form of **free-space electromagnetic waves** which radiate away from the line with the velocity of light.

The 19th century mathematician James Clerk Maxwell was the first to recognise that electromagnetic waves can exist and transport energy quite independently of any system of conductors. We know now that radio waves, heat waves, visible light, X-rays are all electromagnetic waves differing only in frequency. Figure 2.1-9 shows the range of frequencies and the regions occupied by the different types of radiation. This is known as the **electromagnetic spectrum**.

2.1.4.2 Free-space radiation

Introduction

At operational frequencies, where the operational wavelengths are comparable in size to circuit components,[11] any circuit consisting of components connected by

[9] Strictly speaking 'free space' is a vacuum. However, the velocity of propagation of electro-magnetic waves in the atmosphere is practically the same as that in a vacuum and is approximately 3×10^8 metres per second. Wavelength (λ) is defined as the ratio, velocity/frequency.
[10] Transmission lines have impedances because they are constructed from physical components which have resistance, self-inductance, conductance and capacitance.
[11] Generally taken to be the case when the operational wavelength is about 1/20 of the physical size of components.

Hz	Power and audio signals	Radio signals	Microwaves	Infrared and visible	UV X-rays X-rays->
10^0	10^1 10^4	10^5 10^8	10^9 10^{11}	$10^{12} \sim 10^{15}$	$10^{15} \sim 10^{17}$

Fig. 2.1-9 The electromagnetic frequency spectrum.

conductors will tend to act as an imperfect transmission line. As a result, there will always be some loss of r.f. energy by way of radiation. In other words, the circuit will tend to behave like a crude radio transmitter antenna.

It follows that for minimal radiation, components should be small with respect to their operational wavelengths. Conversely, if radiation is desired, then the physical components should be large, approximately 1/4 wavelength for optimum radiation. This is why antennas are physically large in comparison with their operational wavelength.

Energy radiates from an r.f. source or transmitter in all directions. If you imagine a spherical surface surrounding the transmitter, then the interior of the surface would be 'illuminated' with radiated energy, just like the inside of a globular lamp-shade. The illumination is not necessarily uniform, however, since all transmitters are, to some extent, directional.

If the r.f. source is sinusoidal, then the electric and magnetic fields will also be varying sinusoidally at any point in the radiation field. Now it is difficult to depict a propagating electromagnetic field but some of its important properties can be identified. To do this we consider propagation in a particular direction on a straight line connecting a transmitter to a distant receiver as shown in Figure 2.1-10. You will see that this line coincides with the z-direction in Figure 2.1-10. Measurements at the radio receiver would then indicate that the oscillating electric field is acting all in one direction, the x-direction in Figure 2.1-10. The magnetic field is in-phase with the electric field but acts at right-angles to the electric field, in the y-direction. The two fields are thus at right-angles to each other and to the direction of propagation. An electromagnetic wave with these characteristics is known as a **plane wave**.

Polarisation

Provided there is no disturbance in the propagation path, the electric and magnetic field orientations with respect to the earth's surface will remain unchanged. By convention, the orientation of the electric field with respect to the earth's surface is called the **polarisation** of the electromagnetic wave. If the electric field is vertical, the wave is said to be **vertically polarised**; if horizontal, the wave is **horizontally polarised**. A wave is **circularly polarised** if its electric field rotates as the wave travels. Circular polarisation can be either *clockwise* or *anti-clockwise*.

Polarisation is important because antennas must be mounted in the correct plane for optimum signal reception.[12] Terrestrial broadcasting stations tend to use either vertical or horizontal polarisation. Satellite broadcasting stations use circular polarisation. The polarisation of a wave is sometimes 'twisted' as it propagates through space. This twisting is caused by interfering electric or magnetic fields. It is particularly noticeable near steel-structured buildings where aerials are mounted at odd angles to the vertical and horizontal planes to compensate for these effects.

Field strength

The strength of a radio wave can be expressed in terms of the strength of its electric field or by the strength of its magnetic field. You should recall that these are measured in units of volts per metre and amperes per metre respectively. For a sinusoidally varying field it is customary to quote r.m.s. values E_{rms} and H_{rms}. What is the physical significance of E_{rms}? This is numerically equal to the r.m.s. voltage induced in a conductor of length 1 m when a perpendicular electromagnetic wave sweeps over the conductor with the velocity of light.

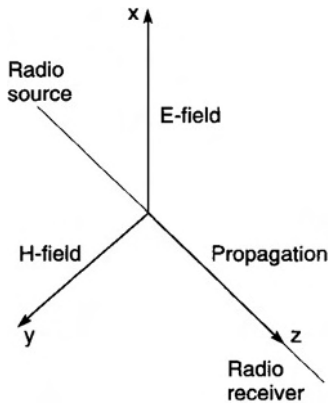


Fig. 2.1-10 Electric and magnetic field directions for an electromagnetic wave propagating in the z-direction.

[12] You can see this effect by looking at TV aerials mounted on houses. In some districts, you will see aerials mounted horizontally whilst in other areas you will find aerials mounted vertically. As a general rule, TV broadcasting authorities favour horizontal polarisation for main stations and vertical polarisation for sub- or relay stations.

As stated earlier, the electric and magnetic fields in a plane wave are everywhere in phase. The ratio of the field strengths is always the same and is given by

$$\frac{\text{electric field strength } E_{\text{rms}}\ (\text{V/m})}{\text{magnetic field strength } H_{\text{rms}}(\text{A/m})} = 377\ \Omega \tag{2.1.8}$$

This ratio is called the free-space **wave impedance**. It is analogous to the characteristic impedance of a transmission line.

Example 2.1.1

The electric field strength at a receiving station is measured and found to have an r.m.s. value of 10 microvolts/m. Calculate (a) the magnetic field strength; (b) the amount of power incident on a receiving aerial with an effective area of 5 m^2.

Given: Electric field strength =10 microvolts/m.
Required: (a) Magnetic field strength, (b) incident power on a receiving aerial with effective area of 5 m^2.

Solution

Using equation 2.1.8

(a) $H_{\text{rms}} = 10\ \mu\text{V/m}/377\ \Omega = 2.65 \times 10^{-8}$ A/m
(b) Power density is given by

$$\begin{aligned} E_{\text{rms}} \times H_{\text{rms}} &= 10 \times 10^{-6} \times 2.65 \\ &\quad \times 10^{-8}\ \text{W/m}^2 \\ &= 2.65 \times 10^{-13}\ \text{W/m}^2 \end{aligned}$$

This is the amount of power incident on a surface of area 1 m^2. For an aerial with area 5 m^2, the total incident power will be

$$P = 2.65 \times 10^{-13}\ \text{W/m}^2 \times 5\ \text{m}^2 = 1.33\ \text{pW}$$

Power density

The product $E_{\text{rms}} \times H_{\text{rms}}$ has the dimensions of 'volts per metre' times 'amps per metre', giving watts per square metre. This is equivalent to the amount of r.f. power flowing through one square metre of area perpendicular to the direction of propagation and is known as the **power density** of the wave. The power density measures the intensity of the 'illumination' falling on a receiving aerial.

A plane wave expands outwards as it travels through space from a point source. As a result, the power density falls off with increasing distance from the source. If you have studied any optics then you will be familiar with the idea that the power density falls off as the *square of the distance* from the source, i.e.

$$\frac{P_{\text{D2}}}{P_{\text{D1}}} = \left[\frac{D_1}{D_2}\right]^2 \tag{2.1.9}$$

where $P_{\text{D1}}, P_{\text{D2}}$ = power densities at distances D_1 and D_2 respectively.

Example 2.1.2

If the data in Example 2.1.1 applies to a receiver located 10 km from the transmitter, what will be the values of E_{rms} and H_{rms} at a distance of 100 km?

Given: Data of Example 2.1.1 applied to a receiver at 10 km from transmitter.
Required: (a) E_{rms} at 100 km, (b) H_{rms} at 100 km.

Solution

Using Equation 2.1.9 at a distance of 100 km, the power density will be reduced by a factor $(10/100)^2 = 0.01$, so power density $= 2.65 \times 10^{-15}$ W/m^2. Now, power density $= E_{\text{rms}} \times H_{\text{rms}}$ and since $H_{\text{rms}} = E_{\text{rms}}/377$ (Equation 2.1.7)

$$\frac{E_{\text{rms}}^2}{377} = 2.65 \times 10^{-15}\ \text{W/m}^2$$

Hence

$$E_{\text{rms}} = \sqrt{2.65 \times 10^{-15} \times 377} = 1\ \mu\text{V/m}$$

and

$$H_{\text{rms}} = 1\ \mu\text{V/m}/377\ \Omega = 2.65 \times 10^{-9}\ \text{A/m}$$

Summary of propagation principles

Several important points have been established in Section 2.1.4.

- R.F. energy is radiated by way of travelling electric and magnetic fields which together constitute an electromagnetic wave propagating in free space with the velocity of light.
- In a plane wave, the electric and magnetic fields vary in phase and act at right-angles to each other. Both fields are at right-angles to the direction of propagation.
- The direction of the electric field determines the polarisation of a plane wave.
- At any point, the ratio of the electric and magnetic fields is the same and equal to the wave impedance. This impedance is 377 Ω approximately.

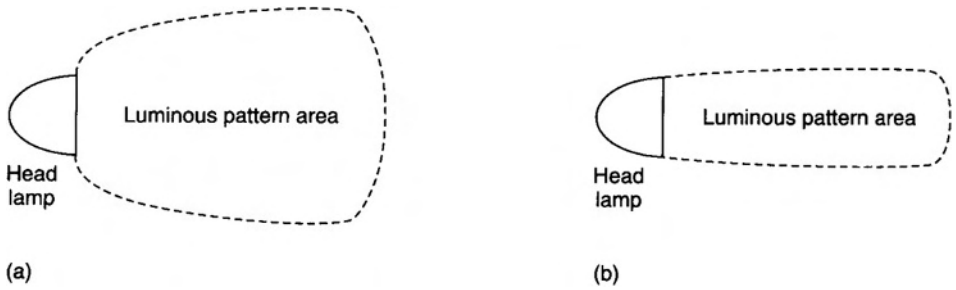


Fig. 2.1-11 Radiation patterns from a car headlamp: (a) top view; (b) side view.

- The product $E_{rms} \times H_{rms}$ gives the power density of the wave.
- The power density falls off as the square of the distance from the r.f. source.
- To obtain optimum signal reception from free space a receiving aerial should be set for the correct polarisation and be suitably located with regard to height and direction.

2.1.5 Antennas and aerials

2.1.5.1 Introduction

An **antenna** or **aerial** is a structure, usually made from good conducting material, that has been designed to have a shape and size so that it will provide an efficient means of transmitting or receiving electromagnetic signals through free space. Many of the principles used in the construction of antennas can be easily understood by analogy to the headlamp of your car (see Figure 2.1-11).

An **isotropic** light source is a light source which radiates light equally in all directions. The radiation pattern from an isotropic light source can be altered by placing a reflecting mirror on one side of the light source. This is carried out in car headlamps where a quasi-parabolic reflecting mirror (reflector) is placed behind a bulb to increase the light intensity of the lamp in the forward direction. The reflector has therefore produced a change in the **directivity** of the light source. The increase or 'gain' of light intensity in the forward direction has been gained at the expense of losing light at the back of the lamp. This gain is not a 'true gain' because total light energy from the lamp has not been increased; light energy has only been *re-directed* to produce an intensity gain in the forward direction.

The forward light intensity of a car lamp can be further improved by using one or more lenses to concentrate its forward light into a main beam or **main lobe**. Again, this 'gain' in light intensity has been achieved by confining the available light into a narrower beam of illumination; there has been no overall gain in light output from the bulb.

There are also optimum sizes and distances for the placement of reflectors and lenses. These are dictated by the physical size of the bulb, the desired gain intensity of the main beam or main lobe, the required width of the main beam and the requirement to suppress minor or spurious light lobes which consume energy and cause unnecessary glare to oncoming motorists.

A car headlamp (Figure 2.1-11) has two main light-emitting patterns: a **horizontal** pattern and a **vertical** pattern. The horizontal pattern (Figure 2.1-11(a)) is a bird's eye view of the illumination pattern. A plot of the horizontal pattern is called a **polar diagram**. The vertical or **azimuth** pattern (Figure 2.1-11(b)) is the pattern seen by an observer standing to one side of the lamp. The vertical pattern is sometimes called the **end-fire** pattern. Both light patterns must be considered because modern headlamp reflectors tend to be elliptical and affect emitted light in the horizontal and vertical planes differently.

In the above description, light has been assumed to travel from bulb to free space but the effect is equally true for light travelling in the opposite direction, i.e. the system is *bi-directional*. It can be used either for *transmitting* light from the bulb or for *receiving* external light at the point source usually occupied by the bulb filament. This can be easily verified by shining an external light source through the lens and the reflector in the opposite direction from which light had emerged, and seeing it converge on the bulb source.[13]

Many of the principles introduced above apply to antennas as well. Because of its bi-directional properties, a radio antenna can be used for transmitting or receiving signals.

[13] If you have any doubts about the system being bi-directional, you should visit a lighthouse which uses a similar reflector and lens system. Curtains must be drawn around the system during daylight hours because sunlight acting on the system has been known to produce such high light and heat intensities that insulation meltdown and fires have been caused.

2.1.5.2 Radiating resistance

The relationship, power (watts) = (volts^2/ohms), is used for calculating power loss in a circuit. It is not always possible to apply this law directly to a radiating circuit because a physical resistor does not always exist. Yet we cannot deny that there is a radiated power loss when a voltage is applied across a radiating circuit. To overcome this problem, engineers postulate an 'equivalent' resistor to represent a physical resistor which would absorb the same radiated power loss. This equivalent resistor is called the **radiating resistance** of the circuit.

The radiating resistance of an antenna should not be confused with its input impedance. The input impedance is the value used when considering the connection of an antenna to a transmission line with a specified characteristic impedance. Antennas are bi-directional and it is not uncommon to use the same antenna for transmitting and receiving signals.

Example 2.1.3

A transmitter with an output resistance of 72 Ω and an r.m.s. output of 100 V is connected via a matched line to an antenna whose input resistance is 72 Ω. Its radiation resistance is also 72 Ω. Assuming that the antenna is 100% efficient at the operating frequency, how much power will be transmitted into free space?

Given: Transmitter output = 100 V, transmitter output impedance = 72 Ω, antenna input impedance = 72 Ω, radiation resistance = 72 Ω, antenna efficiency = 100%.

Required: Power radiated into free space.

Solution

The antenna has an input impedance Z_{in} = 72 Ω and provides a matched termination to the 72 Ω line. The r.f. generator then 'sees' an impedance of 72 Ω, so the r.m.s. voltage applied to the line will be 100/2 = 50 V. The amount of power radiated is calculated using

$$\text{radiated power} = \frac{50^2}{R}$$

where R = 72 Ω is the radiation resistance. The radiated power is therefore 34.7 W. Notice that, because in this case $R = Z_{in}$, maximum power is radiated into free space.

2.1.5.3 The half-wave dipole antenna

Most antennas can be analysed by considering them to be transmission lines whose configurations and physical dimensions have been altered to present easy energy transfer from transmission line to free space. In order to do this effectively, most antennas have physical sizes comparable to their operational wavelengths.

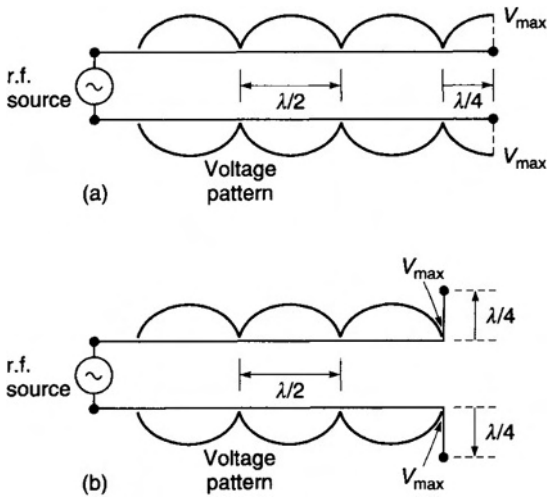


Fig. 2.1-12 (a) Voltage standing-wave pattern on an open-circuited transmission line; (b) open-circuited line forming a dipole.

Figure 2.1-12(a) shows a two-wire transmission line, open-circuited at one end and driven by a sinusoidal r.f. generator. Electromagnetic waves will propagate along the line until it reaches the open-circuit end of the line. At the open-circuit end of the line, the wave will be reflected and travel back towards the sending end. The forward wave and the reflected wave then combine to form a voltage standing wave pattern on the line. The voltage is a maximum at the open end. At a distance of one-quarter wavelength from the end, the voltage standing wave is at a minimum because the sending wave and the reflected wave oppose each other. Suppose now that the wires are folded out from the $\lambda/4$ points, as in Figure 2.1-12(b). The resulting arrangement is called a **half-wave dipole antenna**.

Earlier we said that the electromagnetic fields around the parallel conductors overlap and cancel outside the line. However, the electromagnetic fields along the two ($\lambda/4$) arms of the dipole are now no longer parallel. Hence there is no cancellation of the fields. In fact, the two arms of the dipole now act in series and are additive. They therefore reinforce each other. Near to the dipole the distribution of fields is complicated but at a distance of more than a few wavelengths electric and magnetic fields emerge in phase and at right-angles to each other which propagate as an electromagnetic wave.

Besides being an effective radiator, the dipole antenna is widely used as a VHF and TV receiving antenna. It has a polar diagram which resembles a figure of eight (see Figure 2.1-13). Maximum sensitivity occurs for a signal arriving broadside on to the antenna. In this direction the 'gain' of a dipole is 1.5 times that of an isotropic antenna. An isotropic antenna is a theoretical antenna that radiates or receives signals uniformly in all directions.

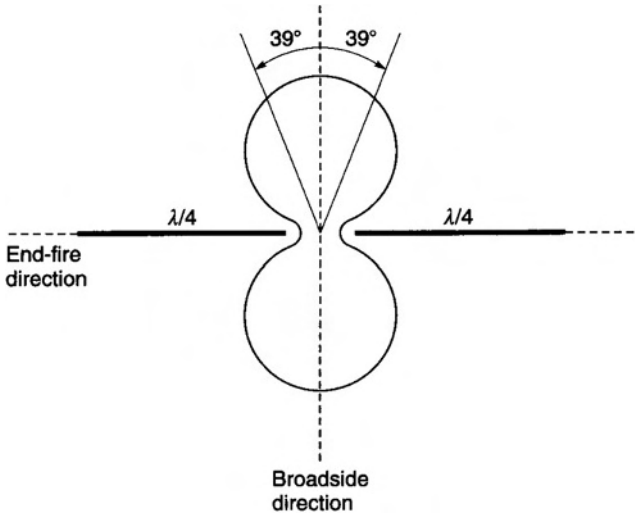


Fig. 2.1-13 Polar pattern of a half-wave dipole.

The gain is a minimum for signals arriving in the 'end-fire' direction. Gain decreases by 3 dB from its maximum value when the received signal is ±39° off the broadside direction. The maximum gain is therefore 1.5 and the half-power beam-width is 78°. The input impedance of a half-wave dipole antenna is about 72 Ω. It turns out that the input impedance and the radiation resistance of a dipole antenna are about the same.

2.1.5.4 Folded dipole antenna

The folded dipole (Figure 2.1-14) is a modified form of the dipole antenna. The antenna is often used for **VHF FM** receivers. The impedance of a folded λ/2 dipole is approximately 292 Ω. This higher input impedance is advantageous for two main reasons:

- it allows easy connection to 300 Ω balanced lines;
- its higher impedance makes it more compatible for use in directive aerials (particularly Yagi arrays); which will be described in Section 2.1.6.

2.1.5.5 The monopole or vertical rod antenna

The monopole or vertical rod antenna (Figure 2.1-15) is basically a coaxial cable[14] whose outer conductor has been removed and connected to earth. It is usually about λ/4 long except in cases where space restrictions or other electrical factors restrict its length. At high frequencies, the required λ/4 length is short and the antenna can be made self-supporting by the use of hollow metal tubing.

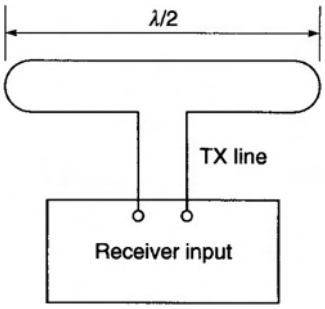


Fig. 2.1-14 Folded dipole antenna.

At low frequencies where a greater length is required, the antenna is often supported by poles.

This antenna is favoured for use in low frequency transmitting stations, in portable radio receivers, in mobile radio-telephones, and for use on motor vehicles because it has a circular polar receiving pattern, i.e. it transmits and receives signals equally well in all directions around its circumference. This is particularly important in mobile radio-phones and in motor vehicles because a motor vehicle may be moving in any direction with respect to a transmitting station. To minimise interference from the engine of the vehicle and for maximum receiving height, rod aerials are frequently mounted on the roofs of vehicles. These aerials are also often mounted at an angle of about 45° to the horizon to enable them to be receptive to both horizontal and vertical polarisation transmissions.

2.1.5.6 Single loop antennas

Another type of antenna which is frequently used for TV reception is the single loop antenna shown in Figure 2.1-16. This loop antenna usually has an electrical length equal to approximately λ/2 at its operating frequency. It is popular with TV manufacturers because it is comparatively cheap and easy to manufacture. The

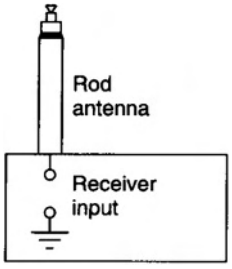


Fig. 2.1-15 Rod or monopole antenna.

[14] A typical example of a coaxial cable is the TV lead which connects your television set to the antenna.

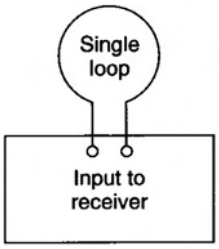


Fig. 2.1-16 Single loop antenna.

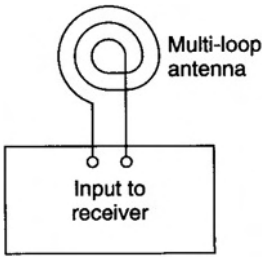


Fig. 2.1-17 Multi-loop antenna.

antenna's input impedance is approximately 292 Ω and it is easily coupled to 300 Ω balanced transmission lines. The antenna is directive and has to be positioned for maximum signal pick-up.

2.1.5.7 Multi-loop antennas

At low frequencies, particularly at frequencies in the medium wave band where wavelengths are long, single loop $\lambda/2$ length antennas are not practical; multi-loop antennas (Figure 2.1-17) are used instead. The multi-loop antenna can be reduced even further in size if a ferrite rod is inserted within the loop.

The open-circuit voltage induced in multiple loop antennas can be calculated by making use of Faraday's Law of Electromagnetic Induction which states that the voltage induced in a coil of n turns is proportional to the rate of change of magnetic flux linkage. For simplicity in derivation, it will be assumed that the incident radiation is propagating along the axis of the coil (see Figure 2.1-18).

Expressing Faraday's Law mathematically,

$$e = n\frac{\mathrm{d}\phi}{\mathrm{d}t} \tag{2.1.10}$$

where

e = open-circuit voltage in volts
n = number of turns on coil
$\mathrm{d}\phi/\mathrm{d}t$ = rate of change of magnetic flux linkage (ϕ = webers and t = seconds)

Some fundamental magnetic relations are also required. These include:

total flux ϕ = flux density (B) per unit area × area (A)

or

$$\phi_{\text{webers}} = B_{\text{tesla}} \times A_{\text{square metres}} \tag{2.1.11}$$

By definition, flux density in air cored coil (B_{tesla}) is given by

free-space permeability (μ_0) × magnetic field strength (H)

or

$$B_{(\text{tesla})} = \mu_{0\ (\text{henry/metre})} \times H_{(\text{ampere/metre})} \tag{2.1.12}$$

or suppose that the incident wave has a magnetic field strength

$$H = H_{\max} \sin \omega t \tag{2.1.13}$$

where ω is the angular frequency of the r.f. signal. Then substituting Equations 2.1.12 and 2.1.13 in Equation 2.1.11 yields

$$\phi = BA = \mu_0 H_{\max} \sin \omega t \times A \tag{2.1.14}$$

Taking the rate-of-change[15] in Equation 2.1.14, then the induced voltage is

$$\begin{aligned} e &= n\frac{\mathrm{d}\phi}{\mathrm{d}t} \\ &= n\omega\mu_0 A H_{\max} \cos \omega t \end{aligned} \tag{2.1.15}$$

For a coil with a ferrite core, the flux density is increased by the relative effective permeability (μ_r), giving

$$e = n\omega\mu_0\mu_r A H_{\max} \cos \omega t \tag{2.1.16}$$

You will see that the ferrite core has increased the effective area of the coil by a factor μ_r. Ferrite cores with effective relative permeabilities of 100–300 are readily available but even with these values, the effective area of the aerial is relatively small when compared with a $\lambda/2$ aerial length. The ferrite rod aerial is therefore very inefficient when compared to an outdoor aerial but it is popular because of its convenient size and portability. At medium wave frequencies, the inherent poor signal

[15] If you do not know how to differentiate to get the rate of change of a value, then please refer to a maths book.

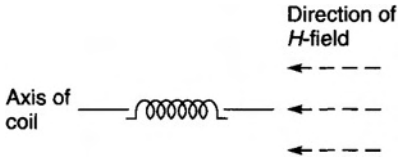


Fig. 2.1-18 Multi-looped antenna aligned for maximum flux linkage.

pick-up is acceptable because broad-cast stations radiate large signals.

In the foregoing derivation, it has been assumed that the magnetic field has been cutting the coil along its axis. Occasions arise when the incident magnetic field arrives at an angle α with respect to the axis of the coil. This is shown in Figure 2.1-19. In this case the effective core area is reduced by cos α, and the induced voltage becomes

$$e = n\omega\mu_0 AH_{\max} \cos \omega t \cos \alpha \qquad (2.1.17)$$

This expression shows that the induced open-circuit voltage, e, is dependent on the axial direction of the aerial coil with respect to the direction of the propagation. It is maximum when cos $\alpha = 1$, i.e. $\alpha = 0°$, and minimum when cos $\alpha = 0$, i.e. $\alpha = 90°$. This explains why it is necessary to position a loop aerial to receive maximum signal from a particular broad-casting station and this is done in a portable radio receiver by orienting its direction.

The above reasons apply equally well to ferrite rod aerials and for these cases we have an induced voltage

$$e = n\omega\mu_r\mu_0 AH_{\max} \cos \omega t \cos \alpha \qquad (2.1.18)$$

If the magnetic field strength is given as an r.m.s. value (H_{rms}), then the r.m.s. value of the induced voltage is

$$e_{\text{rms}} = n\omega\mu_r\mu_0 AH_{\text{rms}} \cos \alpha \qquad (2.1.19)$$

Finally, ferrite aerials are seldom used at the higher frequencies because ferrite can be extremely lossy above 10 MHz.

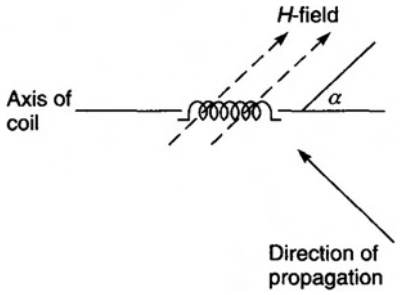


Fig. 2.1-19 *H* field arriving at an angle α.

Example 2.1.4

A coil of 105 turns is wound on a ferrite rod with an effective cross-sectional area of 8×10^{-5} m^2. The relative permeability of the ferrite is 230 and the permeability of air is $4\pi \times 10^{-7}$ henry/m. The r.m.s. field strength is 10 μA/m. If the magnetic field is incident along the axis of the coil and the frequency of operation is 1 MHz, what is the r.m.s. open-circuit voltage induced in the coil?

Given: No. of coil turns =105, effective cross-sectional area of ferrite rod = 8×10^{-5} m^2, relative permeability (μ_r) = 230, permeability of air (μ_0) = $4\pi \times 10^{-7}$ henry/m, r.m.s. field strength =10 μA/m, frequency = 1 MHz.

Required: r.m.s. open-circuit voltage induced in coil.

Solution

Using Equation 2.1.19

$$\begin{aligned} e_{\text{rms}} &= n\omega\mu_r\mu_0 AH_{\text{rms}} \cos \alpha \\ &= 105 \times 2\pi \times 1 \times 10^6 \times 230 \times 4\pi \times 10^{-7} \\ &\quad \times 10 \times 10^{-6} \times 8 \times 10^{-5} \times \cos 0° \\ &= 152.5 \ \mu\text{V} \end{aligned}$$

Broadcasting authorities tend to quote electric field strengths rather than magnetic field strengths for their radiated signals. This creates no problems because the two are related by the wave impedance formula given earlier as Equation 1.8. This is repeated below:

$$\frac{\text{electric field strength } (E)}{\text{magnetic field strength } (H)} = 377\ \Omega$$

Example 2.1.5

A coil of 100 turns is wound on a ferrite rod with an effective cross-sectional area of 8×10^{-5} m^2. The relative permeability of the ferrite is 200 and the permeability of air is $4\pi \times 10^{-7}$ henry/m. The magnetic field is incident at an angle of 60° to the axis of the coil and the frequency of operation is 1 MHz. If the electric field strength is 100 μV/m, what is the r.m.s. open-circuit voltage induced in the coil?

Given: No. of coil turns = 105, effective cross-sectional area of ferrite rod = 8×10^5 m^2, relative permeability (μ_r) = 200, permeability of air (μ_0)=$4\pi \times 10^7$ henry/m, incidence of magnetic field = 60° frequency = 1 MHz, electric field strength = 100 μV/m.

Required: Open-circuit voltage (e_{rms}).

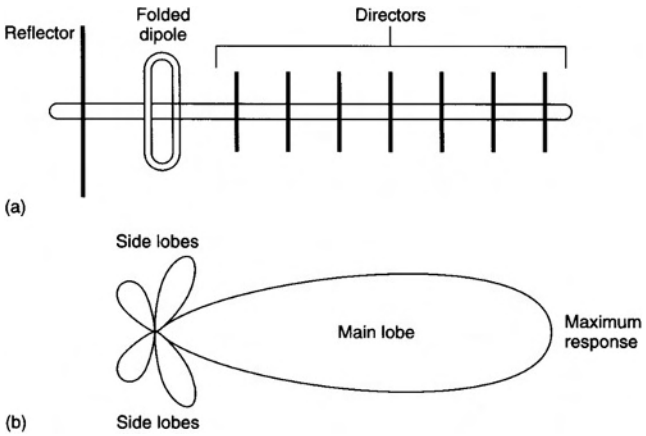


Fig. 2.1-20 Yagi-Uda array: (a) physical arrangement; (b) radiation pattern.

Solution

Substituting Equation 2.1.8 in Equation 2.1.19 yields

$$\begin{aligned} e_{\text{rms}} &= n\omega\mu_r\mu_0 A\frac{E_{\text{rms}}}{377}\cos\alpha \\ &= 100\times 2\pi\times 1\times 10^6\times 200\times 4\pi\times 10^{-7}\times 8 \\ &\quad \times 10^{-5}\times\frac{100\times 10^{-6}}{377}\times\cos 60^\circ \\ &= 1.68\ \mu\text{V} \end{aligned}$$

2.1.6 Antenna arrays

2.1.6.1 Introduction

Antenna arrays are used to shape and concentrate energy in required patterns. One of the more common domestic arrays is the **Yagi-Uda array** used for the reception of television signals.

2.1.6.2 Yagi-Uda array

The Yagi-Uda aerial array shown in Figure 2.1-20 is one of the most commonly used antenna arrays. It is used extensively for the reception of TV signals and can be seen on the roofs of most houses. The Yagi array is an antenna system designed with very similar principles to the car headlamp system described in Section 2.1.5.1. Its main elements are a folded dipole, a reflector, and directivity elements which serve as 'electrical lenses' to concentrate the signal into a more clearly defined beam. The number of directors per array varies according to the gain required from the aerial. The length of directors and the spacing between them are also dependent on the number of elements used in the array. In general, gain increases with the number of directors, but greater gain needs more careful alignment with the transmitting station and requires that the antenna be more sturdily mounted otherwise its pointing direction will waver in high winds which can cause fluctuations in the received signal strength.

Table 2.1-1 Typical values for Yagi arrays operating in the TV band.

No. of elements	Forward gain (±0.5 dB)	Front/back ratio (±2 dB)	Acceptance angle (±3°)
10	12	27	21
13	13	28	19
14	17	30	14
21	19	31	12

The Yagi array is usually designed to be connected to a 75 Ω transmission line.[16] Yagi arrays suitable for operation over the entire TV band can be obtained commercially, but these broadband arrays are usually designed to 'trade off' bandwidth against aerial gain. Broadband Yagi arrays are extremely useful for mobile reception where minimum space and convenience are of importance. (You often see them on top of mobile caravans.) Domestic Yagi arrays are usually designed to provide greater gain but with a more restricted operational frequency band. The latter is not a disadvantage because TV stations operating from a common transmitting site confine their broadcasts to well defined frequency bands. The common practice for domestic Yagi arrays[17] is to use three or more designs (scaled in size) to provide reception for the complete TV band.

Typical values for Yagi arrays operating in the TV band are shown in Table 2.1-1. These figures have been taken from a well known catalogue but some of the terms need explanation.

- 'Number of elements' means the total number of directors, folded dipoles and reflectors used in the array. For example, if the number of elements in an array is 10, the array includes eight directors, one folded dipole and one reflector.

[16] Earlier on, we said that the impedance of a folded dipole aerial was 292 Ω, yet now we say that this antenna is designed to operate with a 75 Ω system. This apparent discrepancy arises because the use of reflector and directors loads the folded dipole and causes its impedance to fall. Judicious director spacing is then used to set the array to the required impedance.

[17] There is a class of Yagi arrays known as Log Periodic Yagis. These have greater bandwidths because the directors are spaced differently. They do cover the entire TV bands but their gain is a compromise between frequency bandwidth and gain.

- 'Forward gain' is the maximum 'gain' which the antenna can provide with respect to an isotropic aerial. A maximum aerial gain of 10 dB means that the antenna will provide 10 times the 'gain' you would get from an isotropic aerial when the array is pointed in its maximum gain direction.
- 'Front to back ratio' is the difference in gain between the direction of maximum antenna gain and the minimum direction of gain which is usually in the opposite direction. This ratio is important because it provides a measure of how the array behaves towards interfering signals arriving from different directions. It is particularly useful in confined areas such as cities where interfering signals 'bounce' off high buildings and interfere with a strong desired signal. In such cases, it is often better to select an antenna with a large front to back ratio to provide rejection to the interfering signal than trying to get maximum antenna gain.
- 'Acceptance angle' is the beamwidth angle in degrees where antenna gain remains within 3 dB of its stated maximum gain. An acceptance angle of 20° and a maximum array gain of 10 dB means that for any signal arriving within ±10° of the maximum gain direction the antenna will provide at least (10–3) dB, i.e. 7 dB of gain. However, you should be aware that the acceptance angle itself is not accurate and that it can vary by ±3° as well.

The values given in the table are representative of the middle range of commercially available Yagi arrays. The figures quoted above have been measured by manufacturers under ideal laboratory conditions and proper installation is essential if the specification is to be achieved in practice.

2.1.7 Antenna distribution systems

Occasions often arise where it is desired to have one antenna supply signal to several television and radio receivers. A typical example is that of an apartment block, where a single aerial on the roof supplies signals to all the apartments. Another possible use for such a system is in your own home where you would like to distribute signals to all rooms from a single external aerial. In such cases, and for maximum efficiency, an aerial distribution system is used. There are many ways of designing such a system but before discussing them, it is best to understand some of the terms used.

2.1.7.1 Balanced and unbalanced systems

Examples of balanced and unbalanced aerials and distribution lines are shown in Figures 2.1.21 and 2.1.22. You should refer to these figures while you are reading the descriptions given below.

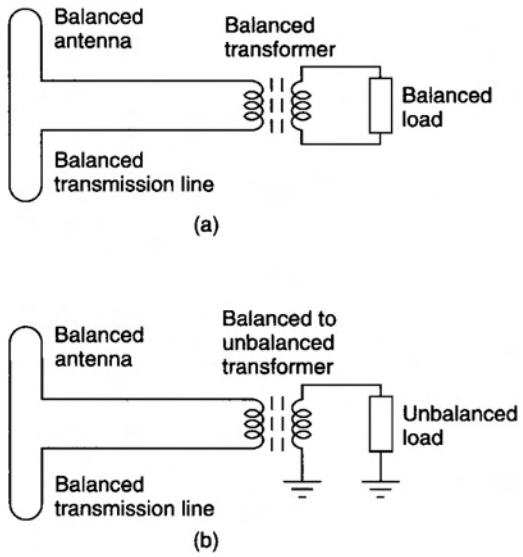


Fig. 2.1-21 Balanced antenna system and (a) balanced distribution system; (b) unbalanced distribution system.

A **balanced antenna** (Figures 2.1-21(a) and (b)) is an aerial which has neither conductor connected directly to earth; it is balanced because the impedance between earth and each conductor is the same. A folded dipole is a typical example of a balanced antenna because the impedance from each end of the antenna to earth is equal and balanced. An **unbalanced antenna** (Figures 2.1-22(a) and (b)) is an aerial which has one of its conductors connected directly to earth. The impedance between earth and each conductor is not the same. A monopole aerial is a typical example of an unbalanced aerial because its other end (see Figure 2.1-15) is connected to earth.

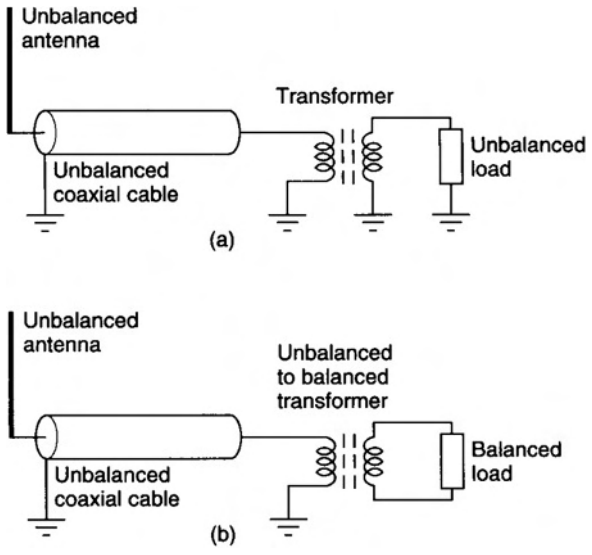


Fig. 2.1-22 Unbalanced antenna system and (a) unbalanced distribution system; (b) balanced distribution system.

A **balanced line** (Figures 2.1-21(a) and (b)) is a transmission line where the impedance between earth and each conductor is identical. A twin pair cable is an example of a balanced line because the impedance between earth and each conductor is the same. An **unbalanced line** (Figures 2.1-22(a) and 2.1-22(b)) is a transmission line where the impedance between earth and each conductor is not equal. A coaxial cable is an example of an unbalanced line because the impedance between earth and the outer shield is different to the impedance between earth and the inner conductor.

The key to the connections in Figures 2.1.21 and 2.1.22 is the **balanced/unbalanced transformer**. These transformers are carefully wound to produce maximum energy transfer by magnetic coupling. Coil windings are designed to have minimum self-capacitance, minimum inter-winding capacitance and minimum capacity coupling between each winding and earth. No direct connection is used between input and output circuits. The above conditions are necessary, otherwise balanced circuits will become unbalanced when parts of the circuit are connected together. The balanced/unbalanced transformer is bi-directional; it can be used to pass energy in either direction.

As the operational frequencies become higher and higher (above 2 GHz), it becomes increasingly difficult to make such a good transformer and a transformer is simply not used and antennas and transmission lines are connected directly. In such cases, the systems resolve to either an unbalanced antenna and distribution system or a balanced antenna and distribution system. The unbalanced system is almost always used because of convenience and costs.

2.1.7.2 Multi-point antenna distribution systems

In the design of antenna distribution systems, transmission lines connecting signal distribution points must function efficiently; they must carry signal with minimum loss, minimum interference and minimum reflections. Minimum loss cables are made by using good conductivity materials such as copper conductors and low loss insulation materials. Minimum interference is obtained by using coaxial cables whose outer conductor shields out interference signals. Reflections in the system are minimised by proper termination of the cables. For proper termination and no reflections in the system, two conditions must be fulfilled:

- the antenna and cable must be terminated in its characteristic impedance Z_0;
- the source impedance (Z_s) feeding each receiver must be matched to the input impedance of the receiver (Z_{in}), i.e. $Z_s = Z_{in}$, otherwise there will be signal reflections and minimum cable transmission loss will not be obtained.

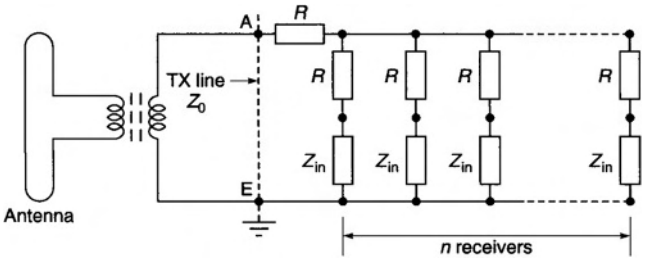


Fig. 2.1-23 Aerial distribution system for *n* receivers, each with an input impedance of Z_{in}.

In Figure 2.1-23, an aerial of characteristic impedance (Z_0) is used to feed a transmission (TX) line with a characteristic impedance Z_0. The output of the line is fed to a number *(n)* of receivers, each of which is assumed to have an input impedance (Z_{in}) equal to Z_0. Resistors R represent the matching network resistors which must be evaluated to ensure properly terminated conditions.

For the system to be properly terminated, it is essential that the aerial and cable system be terminated with Z_0, i.e. the impedance to the fight of the plane 'AE' must present an impedance Z_0 to the antenna and cable system. It is also essential that each receiver be energised from a source impedance (Z_s) matched to its own input impedance (Z_{in}), i.e. $Z_s = Z_{in}$. For ease of analysis we will assume the practical case, $Z_s = Z_{in} = Z_0$.

Now for the transmission line in Figure 2.1-23 to be properly terminated:

$$R+[R + Z_0]/n = Z_0$$

Multiplying both sides by n:

$$nZ_0 = nR + R + Z_0$$

Collecting and transposing terms gives

$$R\frac{(n-1)}{(n+1)}Z_0 \qquad (2.1.20)$$

This equation is all we need to calculate the value of the matching resistors in Figure 2.1-23.

Example 2.1.6

A 75 Ω aerial system is used to supply signals to two receivers. Each receiver has an input impedance of 75 Ω. What is the required value of the matching resistor?

Given: 75 Ω aerial system, input impedance of each receiver = 75 Ω, no. of receivers = 2.

Required: Value of matching resistor.

Solution

Using Equation 2.1.20 with $n = 2$, we obtain

$$R = \frac{(n-1)}{(n+1)}Z_0 = \frac{(2-1)}{(2+1)}75 = 25\ \Omega$$

Example 2.1.7

A 50 Ω aerial receiving system is to be used under matched conditions to supply signal to four receivers, each of input impedance 50 Ω. If the configuration shown in Figure 2.1-23 is used, calculate the value of the resistor, R, which must be used to provide matching conditions.

Given: 50 Ω aerial system, input impedance of each receiver = 50 Ω, no. of receivers = 4.
Required: Value of matching resistor.

Solution

From Equation 2.1.20

$$R = \frac{(n-1)}{(n+1)}Z_0 = \frac{(4-1)}{(4+1)}50 = 30\ \Omega$$

From the answers above, it would appear that an aerial system can be matched to any number of receivers. This is true only within limits because the signal level supplied to individual receivers decreases with the number of distribution points. With large numbers of receivers, network losses become prohibitive.

Transmission losses associated with the matching network of Figure 2.1-23 can be calculated by reference to Figure 2.1-24. The network has been re-drawn for easier derivation of circuit losses but Z_0, R and n still retain their original definitions.

In Figure 2.1-24

V_{oc} = open-circuit source voltage from the aerial
V_{ce} = terminated voltage at an intermediate point in the network
V_{out}= terminated voltage at the input to a receiver

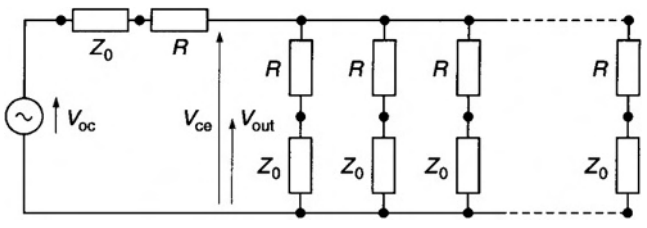


Fig. 2.1-24 Calculating the signal loss in an antenna distribution system.

By inspection

$$V_{out} = \frac{Z_0}{R+Z_0}V_{ce}$$

and

$$V_{ce} = \frac{\left\{\frac{R+Z_0}{n}\right\}}{\left\{\frac{R+Z_0}{n}\right\}+R+Z_0}V_{oc}$$

Therefore

$$V_{out} = \frac{Z_0}{R+Z_0} \times \frac{R+Z_0}{(n+1)(R+Z_0)}V_{oc}$$
$$= \frac{Z_0}{(n+1)(R+Z_0)}V_{oc}$$

Using Equation 2.1.20 and substituting $R = [(n-1)/(n+1)]Z_0$ in the above equation

$$V_{out} = \frac{Z_0}{(n+1)\left[\frac{(n-1)}{(n+1)}Z_0+Z_0\right]}V_{oc} = \frac{V_{oc}}{2n}$$

Transposing, we find that

$$\text{voltage transmission loss} = \frac{V_{out}}{V_{oc}} = \frac{1}{2n} \qquad (2.1.21)$$

or

$$\text{voltage transmission loss} = 20\log[1/2n]\text{dB}^{18}$$

Example 2.1.8

A broadcast signal induces an open-circuit voltage of 100 μV into a rod aerial. The aerial system has a characteristic impedance of 50 Ω and it is used to supply signal to three identical receivers each of which has an input impedance of 50 Ω. If the matching network type shown in Figure 2.1-23 is used, calculate (a) the value of the resistance (R) required for the matching network and (b) the terminated voltage appearing across the input terminals of the receiver.

Given: 50 Ω aerial system, input impedance of each receiver = 50 Ω, no. of receivers = 3, open-circuit voltage in aerial =100 μV.
Required: (a) Value of matching resistor, (b) terminated voltage at receiver input terminal.

Solution

(a) For the matching network of Figure 2.1-23

$$R = \frac{(n-1)}{(n+1)}Z_0 = \frac{(3-1)}{(3+1)} \times 50 = 25\ \Omega$$

(b) Using Equation 2.1.20

$$V_{\text{receiver}} = \frac{1}{2n}V_{\text{antenna}} = \frac{1}{6} \times 100\ \mu\text{V} = 16.67\ \mu\text{V}$$

2.1.7.3 Other aerial distribution systems

The matching network shown in Figure 2.1-23 is only one type of matching network. Figure 2.1-25 shows a commercially available matching network for two outlets. This network is sometimes called a **two-way splitter** because it splits the signal from a single input port into two output ports. The circuit has been designed for low insertion loss and it does this by trading off proper matching against insertion loss.

Example 2.1.9

Figure 2.1-25 shows a commercially available 75 Ω matching network. Calculate: (a) the ratio $V_{\text{out}}/V_{\text{oc}}$ when all ports are each terminated with 75 Ω, (b) the input impedance to the matching network when the output ports are each terminated with 75 Ω and (c) the source impedance to either receiver when the remaining ports are each terminated in 75 Ω.

Given: 75 Ω network splitter of Figure 2.1-25 with 75 Ω terminations.

Required: (a) Ratio $V_{\text{out}}/V_{\text{oc}}$, (b) input impedance of matching network, (c) source impedance to either receiver.

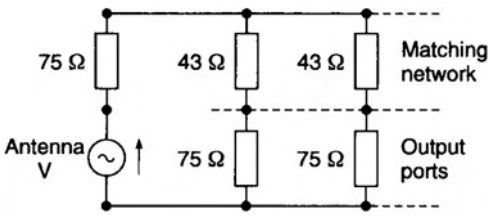


Fig. 2.1-25 Two-way splitter.

Solution

By inspection:

(a) $$V_{\text{out}} = \frac{75}{43+75} \times \frac{(43+75)/2}{(43+75)/2+75} \times V_{\text{oc}} = 0.28$$

(b) input impedance to the network $= (43 + 75)/2 = 59\ \Omega$

(c) receiver source impedance

$$= 43 + \frac{(43+75)(75)}{(43+75)+(75)} = 89\ \Omega$$

From the answers to Example 2.1.9, it can be seen that the insertion loss is slightly reduced but this has been carried out at the expense of system match. The manufacturer is fully aware of this but relies on the fact that the reflected signal will be weak and that it will not seriously affect signal quality. The manufacturer also hopes that the cable system will be correctly matched by the antenna and that any reflections set up at the receiver end will be absorbed by the hopefully matched antenna termination to the cable system. This design is popular because the installation cost of an additional resistor is saved by the manufacturer.

2.1.7.4 Amplified antenna distribution systems

Amplified aerial distribution systems are aerial distribution systems which incorporate amplifiers to compensate for signal transmission, distribution and matching losses. Two systems will be discussed here. The first concerns a relatively simple distribution system where indoor amplifiers are used. The second system deals with a more elaborate system using amplifiers mounted on the aerial (masthead amplifiers) to compensate for distribution and matching losses.

2.1.7.5 Amplified aerial distribution systems using amplifiers

The block diagram of an amplified aerial distribution system using an amplifier is shown in Figure 2.1-26. This system is often used in domestic environments. Outdoor

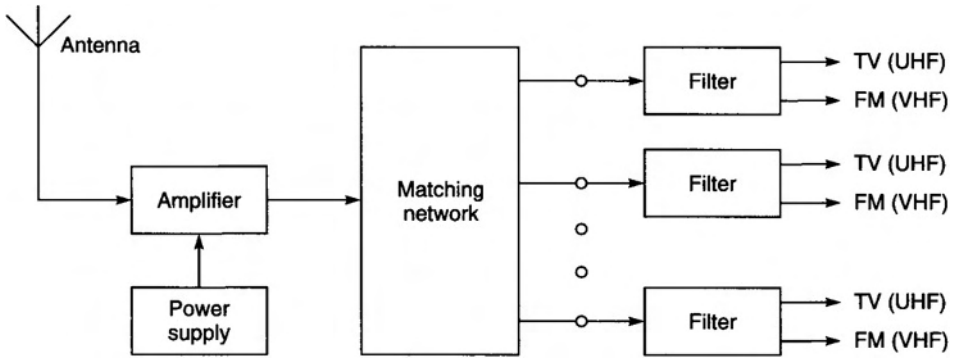


Fig. 2.1-26 Antenna amplifier distribution network.

aerials provide the incoming signals, UHF for TV and/or VHF for FM-radio to the input of an amplifier. The gain of this amplifier is nominally greater than l0 dB but this varies according to the particular amplifier used. The amplifier is usually placed on the antenna masthead or near the aerial down-lead cables and a power supply point in the attic.

Output signals from the amplifier are fed into matching networks for distribution to individual terminals. To save cabling costs, both UHF and VHF signals are often carried on the same cables. Filters (a high pass filter for UHF and a low pass filter for VHF) are installed at individual terminals to feed the signals to their designated terminals.

The main advantage of such a system is that it is relatively easy to install especially if wiring for the distribution already exists. It also compensates for signal loss in the distribution network. The amplifier casing is relatively cheap when the amplifier is used indoors as it does not have to be protected from extreme weather conditions. The main disadvantage of indoor mounting is that signals are attenuated by the aerial down-lead cables before amplification. This signal loss decreases the available signal before amplification and therefore a poorer signal-to-noise ratio is available to the distribution points than if the amplifier was to be mounted on the masthead.

2.1.8 Radio receivers

2.1.8.1 Aims

The aims of this section are to introduce you to:

- the tuned r.f. receiver
- the superhet receiver
- the double superhet receiver
- selectivity requirements in receivers
- sensitivity requirements in receivers
- concepts of signal-to-noise and sinad ratios
- noise figures of receivers

2.1.8.2 Objectives

After reading this section you should be able to understand:

- the basic principles of tuned r.f. receivers (TRFs)
- the basic principles of superhet receivers
- the basic principles of satellite receivers
- the concepts of selectivity
- the concepts of sensitivity
- the concepts of signal-to-noise and sinad ratios
- the concepts of noise figures

2.1.8.3 Introduction

Radio receivers are important because they provide a valuable link in communications and entertainment. Early receivers were insensitive, inefficient, cumbersome, and required large power supplies. Modem designs using as little as one integrated circuit have overcome most of these disadvantages and relatively inexpensive receivers are readily available.

2.1.8.4 Fundamental radio receiver requirements

In the AM medium wave band, broadcasting stations transmit their signals centred on assigned carrier frequencies. These carrier frequencies are spaced 9 kHz apart from each other as in Figure 2.1-27 and range from 522 kHz to 1620 kHz. The information bandwidth allocated for each AM transmission is 9 kHz. This means that modulation frequencies greater than 4.5 kHz are not normally used. To receive information from a broadcast signal, an AM broadcast receiver must be tuned to the correct carrier frequency, have a bandwidth that will pass the required modulated signal, and be capable of extracting information from the required radio signal to operate desired output devices such as loudspeakers and earphones.

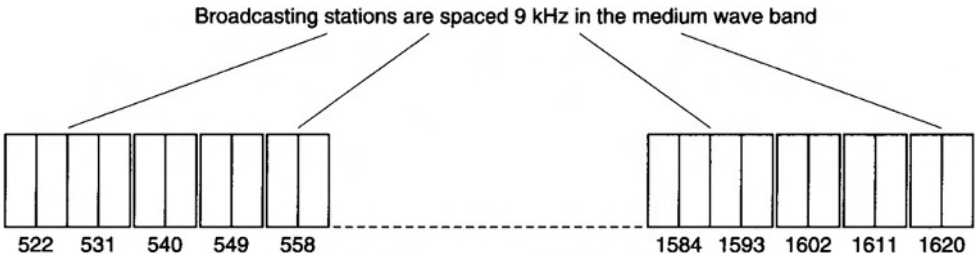


Fig 2.1-27 Spacing of broadcast stations in the medium wave band.

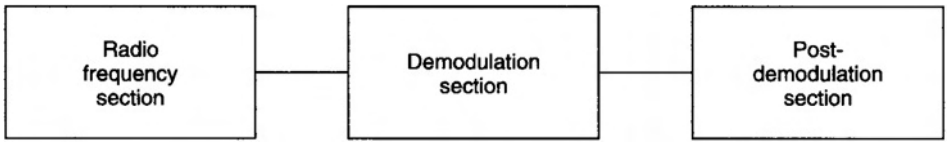


Fig. 2.1-28 Three main sections of a radio receiver.

Discussions that follow pertain mainly to receivers operating in this band. This is not a limitation because many of the principles involved apply equally well to other frequency bands. When the need arises, specific principles applying to a particular frequency band will be mentioned but these occasions will be clearly indicated.

2.1.9 Radio receiver properties

A radio receiver has three main sections (see Figure 2.1-28).

- A **r.f. section** to *select* and if necessary to *amplify* a desired r.f. signal to an output level sufficient to operate a demodulator.
- A **demodulator section** to *demodulate* the required radio signal and *extract* its modulated information.
- A **post-demodulation section** to amplify demodulated signals to the required level to *operate* output devices such as loudspeakers, earphones and/or TV screens.

2.1.9.1 Radio frequency section

A r.f. section is designed to have the following properties.

Selectivity

Receiver selectivity is a measure of the ability of a radio receiver to select the desired transmitted signal from other broadcast signals. An ideal selectivity response curve for an AM broadcast receiver centred on a desired carrier frequency (f_0) is shown in Figure 2.1-29. Two main points should be noted about the ideal selectivity response curve. First, it should have a wide enough passband (9 kHz approx.) to pass the entire frequency spectrum of the desired broadcast signal. Second, the passband should present equal transmission characteristics to all frequencies of the desired broadcast signal. In addition, the bandwidth should be no wider than that required for the desired signal because any additional bandwidth will allow extraneous signals and noise from adjacent channels to impinge on the receiver. Notice that the skirts of the ideal selectivity curve are vertical, so that the attenuation of any signal outside the passband is infinitely high.

In practice, costs and stability constraints prevent the ideal selectivity response curve from ever being attained and it is more rewarding to examine what is achieved by commercial receivers.

An overall receiver selectivity response curve for a typical domestic transistor receiver for the reception of AM broadcast signals is shown in Figure 2.1-30. In the table supplied with Figure 2.1-30, you should note that the selectivity curve is not symmetrical about its centre frequency. This is true of most tuned circuits because the effective working quality factor (Q_w) of components,

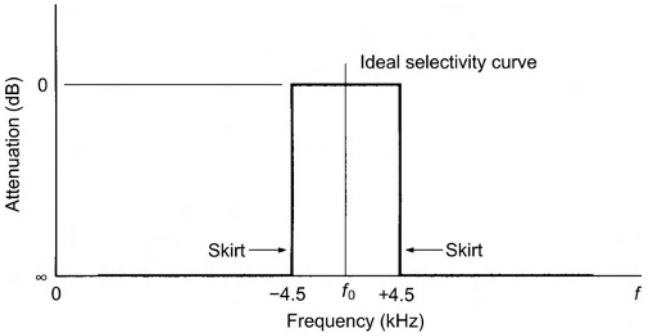
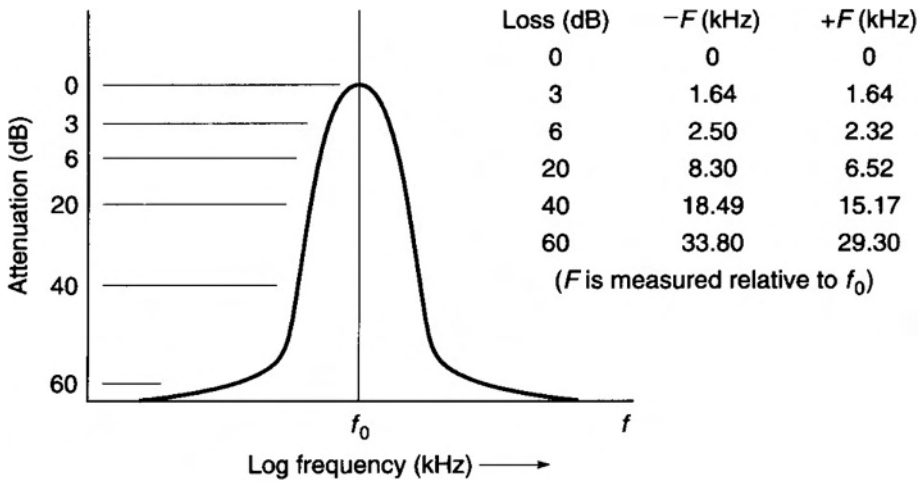


Fig. 2.1-29 Ideal selectivity curve for an AM medium wave broadcast receiver.



Loss (dB)	$-F$ (kHz)	$+F$ (kHz)
0	0	0
3	1.64	1.64
6	2.50	2.32
20	8.30	6.52
40	18.49	15.17
60	33.80	29.30

(F is measured relative to f_0)

Fig. 2.1-30 Typical selectivity curve of a commercial AM six transistor receiver.

particularly inductors, varies with frequency. Note also that the 3 dB bandwidth is only 3.28 kHz and that the 6 dB bandwidth points are approximately 4.82 kHz apart. The 60 dB points are 63.1 kHz apart.

Consider the case of a carrier signal (f_0) modulated with two inner sideband frequencies f_{1L}, f_{1U} ($\pm$1.64 kHz) and two outer sideband frequencies f_{2L}, f_{2U} ($\pm$2.4 kHz) away from the carrier. The frequency spectrum of this signal is shown in Figure 2.1-31(a). When this frequency spectrum is passed through a receiver with the selectivity response shown in Figure 2.1-30, the inner sidebands f_{1L}, f_{1U} ($\pm$1.64 kHz) and the outer sidebands f_{2L}, f_{2U} ($\pm$2.4 kHz) will suffer attenuations of 3 dB and 6 dB approximately with respect to the carrier. (See the table in Figure 2.1-30.) The new spectrum of the signal is shown in Figure 2.1-31(b).

Comparison of Figures 2.1.31(a) and (b) shows clearly that amplitude distortion of the sidebands has occurred but what does this mean in practice? If the transmitted signal had been music, there would have been an amplitude reduction in high notes. If the transmitted signal had been speech, the speaker's voice would sound less natural.

From the above discussion, it should be noted that for good quality reproduction the selectivity curve of a receiver should be wide enough to pass all modulation frequencies without discriminatory frequency attenuation.

Adjacent channel selectivity

A graphical comparison of the selectivity curves of Figures 2.1-29 and 2.1-30 is shown in Figure 2.1-32. From these curves, it can be seen that the practical selectivity curve does not provide complete rejection of signals to stations broadcasting on either side of the desired response channel. The breakthrough of signal from adjacent channels into the desired channel is known as **adjacent channel interference**.

Adjacent channel interference causes signals from adjacent channels to be heard in the desired channel. It is particularly bad when strong adjacent channel signals are present relative to the desired station. What does this mean in practice? It means that you will obtain interference from the unwanted station. In a broadcast receiver, you often hear signals from both channels simultaneously.

Broadcasting authorities minimise adjacent channel interference by forbidding other transmitters situated near the desired station to broadcast on an adjacent channel. Stations geographically distant from the desired

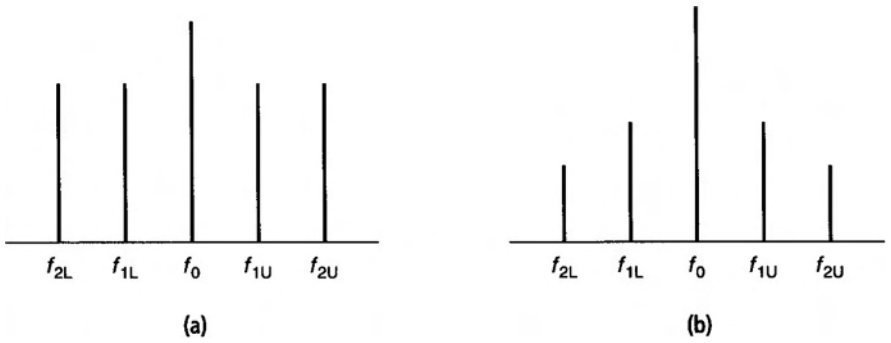


Fig. 2.1-31 (a) Transmitted and (b) distorted spectrum.

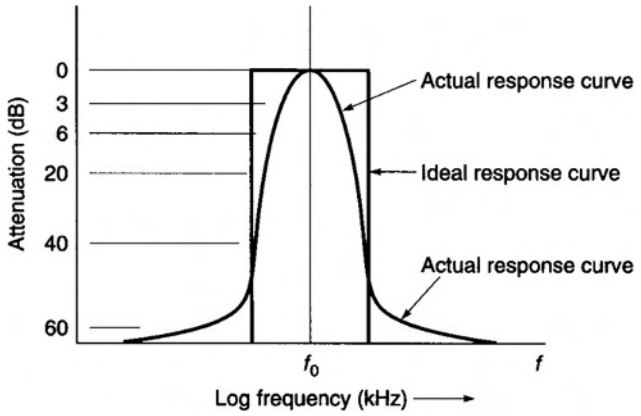


Fig. 2.1-32 A comparison between the ideal and practical selectivity curve.

station are allowed to operate on adjacent channels because it is likely that their signals will have suffered considerable transmission loss by the time they impinge on the desired channel.

Sensitivity

The sensitivity of a radio receiver is a measure of the modulated signal input level which is required to produce a given output level. A receiver with good sensitivity requires a smaller input signal than a receiver with poor sensitivity to produce a given output level. The sensitivity of a small portable receiver (audio output rated at 250 mW) may be quoted as 200 μV/m. What this means is that a modulated AM carrier (modulated with a 400 Hz tone and with an AM modulation depth of 30%) will produce an audio output of 50 mW under its maximum gain conditions when the input signal is 200 μV/m.

2.1.9.2 Signal-to-noise ratios

Any signal transmitted through a communications system suffers attenuation in the passive (non-amplifying) parts of the system. This is particularly true for radio signals propagating between transmitting and receiving aerials. Attenuation is compensated for by subsequent amplification, but amplifiers add their own inherent internally generated random noise to the signal. Noise levels must always be less than the required signal, otherwise the required signal will be lost in noise. Some means must be provided to specify the level of the signal above noise. This means is called the **signal-to-noise ratio**. It is defined as

$$\text{signal-to-noise ratio} = \frac{S}{N} = \frac{\text{signal power}}{\text{noise power}} \qquad (2.1.23)$$

Notice that *S/N* is specified as a ratio of power levels.

An alternative way of specifying signal-to-noise ratios is to quote the ratio in decibels. This is defined by Equation 2.1.24:

$$\text{signal-to-noise ratio} = \frac{S}{N}(\text{dB}) = 10 \log_{10}\left[\frac{\text{signal power}}{\text{noise power}}\right]\text{dB} \qquad (2.1.24)$$

A strong signal relative to the noise at the receiver input is essential for good reception. In practice, we require an *S/N* of 10–20 dB to distinguish speech, an *S/N* of 30 dB to hear speech clearly, and an *S/N* of 40 dB or better for good television pictures.

Noise figure

Certain amplifiers have more inherent electrical noise than others. Manufacturers usually produce a batch of transistors, then classify and name the transistors according to their inherent electrical noise levels. The inherent noise produced by a transistor is dependent on its general operating conditions, particularly frequency, temperature, voltage and operating current, and these conditions must be specified when its noise level is measured. Engineers use the ratio term **noise figure** to specify noise levels in transistors.

Noise figure is defined as

$$\text{noise figure (N.F.)} = \left[\frac{(S/N)_{\text{in}}}{(S/N)_{\text{out}}}\right] \text{at 290 K} \qquad (2.1.25)$$

If a transistor introduces no noise, then its *S/N* at both the input and output is the same; therefore, from Equation 2.1.25, N.F. = 1 or in dB N.F. = 10 log 1 = 0 dB. Hence a 'perfect' or 'noiseless' amplifier has a noise figure of 0 dB. An imperfect amplifier has a noise figure greater than 0 dB. For example an amplifier with a noise figure of 3 dB (= 2 ratio) means that it is twice as bad as a perfect amplifier.

2.1.10 Types of receivers

There are many types of radio receivers. These include:

- tuned TRF
- superheterodyne receivers (superhets)
- double superheterodyne receivers (double superhets)

2.1.10.1 Tuned radio frequency receiver

A tuned r.f. receiver (Figure 2.1-33) has three main sections: a r.f. amplifier section, a detector section, and an audio amplifier section.

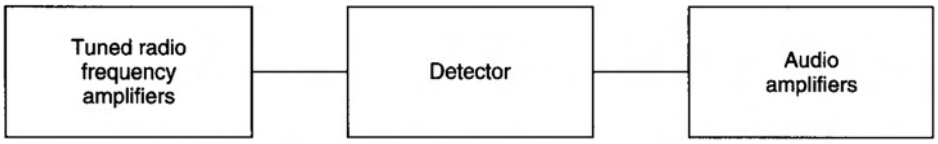


Fig. 2.1-33 Main sections of a tuned radio frequency receiver.

The r.f. section consists of one or more r.f. amplifiers connected in cascade.[19] For efficient operation, all tuned circuit amplifiers must be tuned to exactly the same broadcast frequency and to ensure that this is the case, all tuning adjusters are fixed on to a common tuning shaft. Tuning capacitors which are connected in this manner are said to be 'ganged' and two- and three-stage ganged tuning capacitors are common.

The detector is usually a conventional AM diode type detector. This type of detector is usually a diode which detects the positive peaks of the modulated carrier and filters the r.f. out, so that the remaining signal is the inital low frequency modulation frequency.

The audio section uses audio amplifiers which serve to amplify the signals to operate a loudspeaker. This section is similar to the amplifier in your home which is used for playing compact disks (CDs) and cassettes.

Advantages

The main advantages of TRF receivers are that they are relatively simple, easy to construct, and require a minimum of components. A complete TRF receiver can be constructed using a single integrated circuit such as a ZN414 type chip.

Disadvantages

TRF receivers suffer from two main disadvantages, gain/bandwidth variations and poor selectivity. The inevitable change of gain and bandwidth as the receiver is tuned through its frequency range is due to changes in the selectivity circuits.

Circuit instability can be a problem because it is relatively easy for any stray or leaked signal to be picked up by one of the many r.f. amplifiers in the receiver. R.F. signal can also be easily coupled from one r.f. stage to another through the common power supply. To minimise these risks, r.f. amplifiers are usually shielded and de-coupled from the common power supply.

2.1.10.2 Superheterodyne receiver

Block diagram

A block diagram of a superheterodyne (commonly called superhet) receiver is shown in Figure 2.1-34.

This receiver features an r.f. section which selects the desired signal frequency (f_{rf}). This signal is then mixed with a local carrier at frequency (f_o) in a frequency changer to produce an intermediate frequency (f_{if}) which retains the modulated information initially carried by $f_{rf.}$ The intermediate frequency (f_{if}) then undergoes intensive amplification (60–80 dB) in the intermediate frequency (i.f.) amplifiers to bring the signal up to a suitable level for detection and subsequent application to the post-detection (audio) amplifiers.

R.F. amplifiers are sometimes included in the r.f. section in order to make the noise figure of the receiver as small as possible. Frequency changers have comparatively larger noise figures (6–12 dB) than r.f. amplifiers.

The frequency of the local oscillator (f_o) is always set so that its frequency differs from the desired frequency (f_{rf}) by an amount equal to the intermediate frequency (f_{if}), i.e.

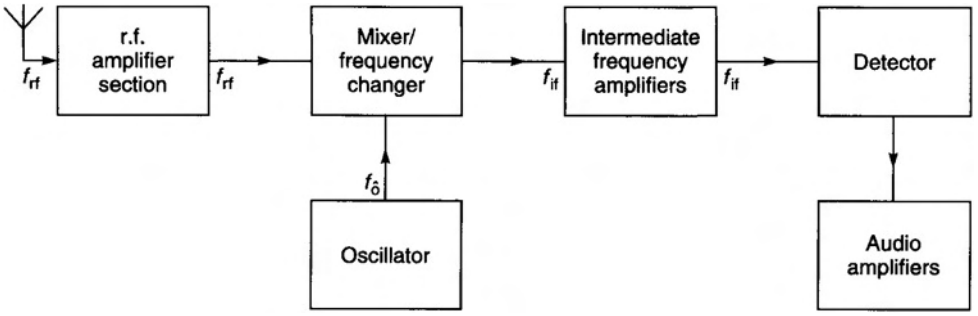


Fig. 2.1-34 Block diagram of a superhet radio receiver.

[19] Here, cascade is meant to imply one amplifier following another amplifer and so on.

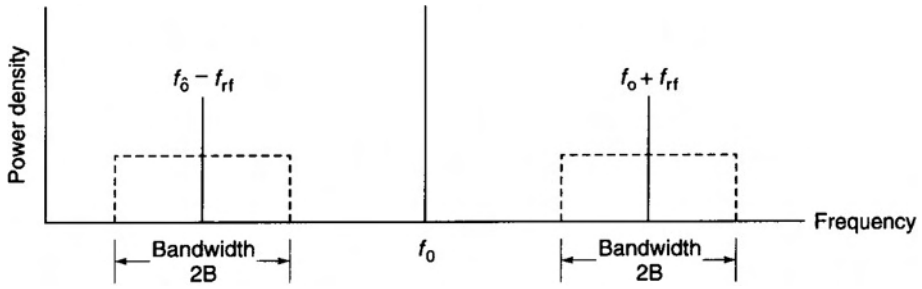


Fig. 2.1-35 Image response in superhet receivers.

$$f_o - f_{rf} = f_{if} \tag{2.1.26}$$

or

$$f_{rf} - f_o = f_{if} \tag{2.1.27}$$

Equation 2.1.26 is more usual for medium-wave receivers. Typical tuning ranges for a medium-wave receiver with $f_{if} = 465$ kHz are 522–1620 kHz for f_{rf} and 987–2085 kHz for f_o.

Advantages

The main advantages of the superhet receiver are as follows.

- Better selectivity because fixed bandpass filters with well defined cut-off frequency points can be used in the i.f. stages of a superhet. Filters and tuned circuits are also less complex because they need to only operate at one frequency, namely the intermediate frequency.
- In a superhet, tuning is relatively simple. A two-ganged capacitor can be used to tune the r.f. and oscillator sections simultaneously to produce the intermediate frequency for the i.f. amplifiers.
- R.F. circuit bandwidths are not critical because receiver selectivity is mainly determined by the i.f. amplifiers.

Disadvantages

The main disadvantages of superhets are as follows.

- **Image channel interference** is caused by the local oscillator (f_o) combining with an undesired frequency (f_{im}) which is separated from the desired frequency (f_{rf}) by twice the intermediate frequency (f_{if}). Expressed mathematically

$$f_{im} = f_{rf} \pm 2f_{if} \tag{2.1.28}$$

The term **2nd channel interference** is another name for image channel interference.

Image channel interference is more easily understood by substituting some arbitrary values into Equations 2.1.26 and 2.1.27. For example, assume that the local oscillator of a superhet is set to 996 kHz and that its i.f. amplifiers operate at 465 kHz. Then, either of two input frequencies, $996 - 465 = 531$ kHz (Equation 2.1.26) or $996 + 465 = 1461$ kHz (Equation 2.1.27), will mix with

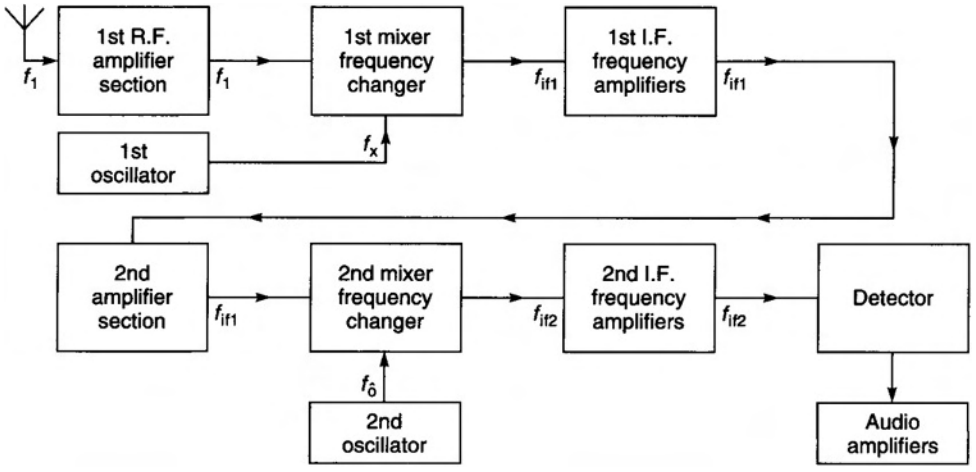


Fig. 2.1-36 Block diagram of a double conversion superhet receiver.

the local oscillator to produce a signal in the i.f. amplifiers. If the desired frequency is 531 kHz, then the undesired frequency of 1461 kHz is 2(f_{if}) or 930 kHz away, i.e. it forms an image on the other side of the oscillator frequency. This condition is shown graphically in Figure 2.1-35.

- There is the possibility that any strong signal or subharmonics of 465 kHz (f_{if}) might impinge directly on the i.f. amplifiers and cause interference.
- Any harmonic of the oscillator (f_o) could mix with an unwanted signal to produce unwanted responses. For example

$$2 \times 996(f_o)\ \text{kHz} - 1527\ \text{kHz} = 465\ \text{kHz}$$

The spurious responses stated above are minimised in superhets by using tuned circuits in the r.f. section of the receiver to select the desired signal and to reject the undesired ones. The local oscillator is also designed to be 'harmonic free'.

2.1.10.3 Double superheterodyne receivers

A block diagram of a double conversion superhet used for receiving direct broadcast signals (DBSs) from satellites is shown in Figure 2.1-36. Direct broadcasting satellites for the United Kingdom region transmit in the 11.6–12.4 GHz band. Each TV channel uses a 26 MHz bandwidth.

The double superhet is basically a superhet receiver with two i.f. sections. The first i.f. section operates at a much higher i.f. frequency than the second i.f. section. This choice is deliberate because a higher 1st intermediate frequency gives better image channel rejection. The 2nd i.f. section is made to operate at a lower frequency because it gives better adjacent channel selectivity.

In this receiver, the input signal, f_1, is selected, mixed with a local oscillator carrier, f_x, and frequency translated to form the first intermediate frequency (f_{if1}). This signal is applied to the 1st i.f. amplifier section, then mixed with f_o to produce a second intermediate frequency (f_{if2}) and amplified prior to detection and low frequency amplification.

In a typical direct broadcast satellite receiver, the first r.f. amplifier section operates in the band 11.6–12.4 GHz. The first local oscillator (f_x) is operated at a fixed frequency of 10.650 GHz. The resultant first i.f. bandwidth range is 950 to 1750 MHz and is really the r.f. band translated to a lower frequency band. This intermediate frequency is then amplified by the first set of 1st i.f. amplifiers. All the foregoing action takes place in a masthead unit which is mounted directly on the antenna. The total gain including r.f. amplification, frequency conversion and i.f. amplification is about 55 dB. This high order of gain is necessary to compensate for the losses which occur in the down-lead coaxial cable to the satellite receiver which is situated within the domestic environment. The satellite receiver treats the 1st i.f. frequency band (950–1750 MHz) as a tuning band and f_o is varied to select the required TV channel which is amplified by the 2nd i.f. section before signal processing.

2.1.11 Summary

The main purpose of this chapter has been to introduce you to the radio environment in your home. The knowledge you have gained will assist you in understanding basic radio propagation and reception principles. It will also help you to remedy some of the simpler radio and TV problems which you are likely to encounter in your home.

In Sections 2.1.1–2.1.3, we started with the necessity for modulation and demodulation and you were introduced to the basic principles of modulation, demodulation and radio propagation. You should now understand the meaning of terms such as **AM**, **FM**, **PM** and digital modulation.

In Section 2.1.4, you were introduced to radio propagation, wave polarisation, field strength and power density of radio waves.

In Sections 2.1.5 and 2.1.6, you learned about the properties of several antennas. These included $\lambda/2$ dipole, folded dipole, monopole, loop antennas and the Yagi-Uda array.

Section 2.1.7 dealt with various antenna distributions and matching systems.

In Section 2.1.8, you encountered some basic concepts concerning the reception of radio signals. You should now be able to carry out simple calculations with regard to selectivity, adjacent channel selectivity, sensitivity, *S/N* ratio and noise figure ratio as applied to radio receivers.

Sections 2.1.9 and 2.1.10 described the main functions required in a radio receiver, and also the main advantages and disadvantages of three basic radio receiver types, namely the TRF, superhet and double superhet receivers. The first type is used in very simple receivers, the second type is used extensively in domestic receivers and the last type is used for direct broadcast reception from satellites.

You have now been provided with an overview of a basic radio communication system.

This page intentionally left blank

Chapter 2.2

2.2

Transmission lines

Ed da Silva

2.2.1 Introduction

At this stage, I would like to prepare you for the use of the software program called PUFF. PUFF (Version 2.2.1) is very useful for matching circuits, and the design of couplers, filters, line transformers, amplifiers and oscillators. Figure 2.2-1 shows what you see when you first open the PUFF program. Figure 2.2-2 shows you how the program can be used in the design of a filter. In Figure 2.2-1 you can see for yourself that to understand and use the program, you must be familiar with Smith charts (top right hand corner) and scattering or '*s*-parameters' (top left hand corner), transmission lines and the methods of entering data (F3 box) into the program. Within limits, the layout window (F1 box) helps to layout your circuit for etching.

2.2.1.1 Aims

We shall cover the basic principles of transmission lines in this part, and Smith charts and *s*-parameters in Section 3. We will then be in a position to save ourselves much work and avoid most of the tedious mathematical calculations involved with radio and microwave engineering.

The main aims of this chapter are:

- to introduce you to various types of transmission lines;
- to explain their characteristic impedances from physical parameters;
- to provide and also to derive expressions for their characteristic impedances;
- to explain their effects on signal transmission from physical and electrical parameters;
- to explain and derive expressions for reflection coefficients;
- to explain and derive expressions for standing wave ratios;
- to explain and derive the propagation characteristics of transmission lines;
- to provide an understanding of signal distortion, phase velocity and group delay;
- to show how transmission lines can be used as inductors;
- to show how transmission lines can be used as capacitors;
- to show how transmission lines can be used as transformers.

2.2.1.2 Objectives

This part is mainly devoted to transmission lines. Knowledge of transmission lines is necessary in order to understand how high frequency engineering signals can be efficiently moved from one location to another. For example, the antenna for your domestic TV receiver is usually mounted on the roof and it is therefore necessary to find some means of efficiently transferring the received signals into your house. In the commercial world, it is not unusual for a radio transmitter to be situated several hundred metres from a mast-mounted transmitting antenna. Here again, we must ensure that minimal loss occurs when the signal is transferred to the antenna for propagation.

2.2.2 Transmission line basics

2.2.2.1 Introduction to transmission lines

In this discussion we shall start off using some basic terms which are easily understood with sound waves. We

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2001, Ed da Silva. Published by Elsevier Ltd 2001. All rights reserved

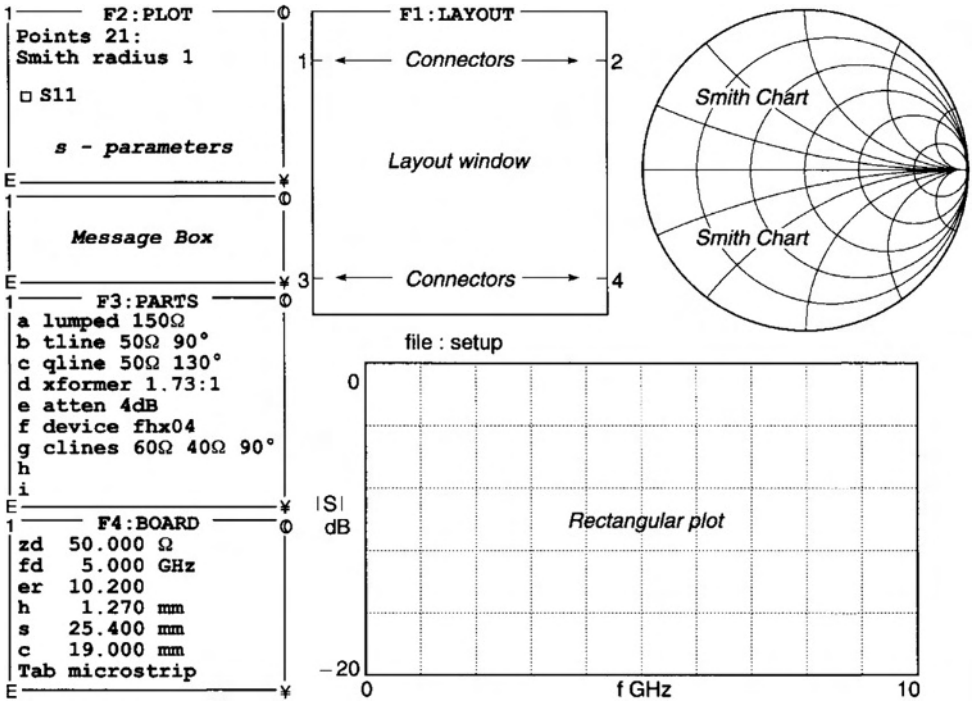


Fig. 2.2-1 PUFF 2.1 – blank screen (words in italics have been added for explanation).

will then use these terms to show that these properties are also applicable to electrical transmission systems. Much of the explanation given in these sections will be based on examples using sinusoids because they are easier to understand. But this information applies equally well to digital waveforms because digital signals are composed of sinusoid components combined in a precise amplitude and phase manner. Therefore, it is vitally important that you do not form the mistaken idea that transmission line theory only applies to analog waveforms.

2.2.2.2 General properties of transmission systems

Transmission systems are used to transfer energy from one point to another. The energy transferred may be sound power or electrical power, or digital/analog/optical signals or any combination[1] of the above.

One easy way of refreshing your memory about signal transmission is to imagine that you are looking into a deep long straight tunnel with walls on either side of you. When you speak, you **propagate** sound energy along a **transmission path** down the length of the tunnel. Your voice is restricted to propagation along the length of the tunnel because walls on either side act as **waveguides.**

Waves emerging directly from the sender are known as **incident waves.** As your vocal cords try to propagate incident waves along the tunnel, they encounter an opposition or **impedance** caused by the air mass in the tunnel. The impedance is determined by the physical characteristics of the tunnel such as its width and height and the manner in which it impedes air mass movement within the tunnel. This impedance is therefore called the **characteristic impedance** (Z_0) of the tunnel.

Bends or rock protrusions along the tunnel walls cause a change in the effective dimensions of the tunnel. These **discontinuities** in effective dimensions can cause minor **reflections** in the signal propagation path. They also affect the characteristic impedance of the transmission channel.

You should note that the walls of the tunnel do not take part in the main propagation of sound waves. However, they do absorb some energy and therefore weaken or **attenuate** the propagated sound energy. Amplitude attenuation per unit length is usually represented by the symbol α.

Moss, lichen and shrubs growing on the walls will tend to absorb high frequency sound better than low frequency sound, therefore your voice will also suffer

[1] For example, the coaxial cable connecting the domestic satellite receiver to the low noise amplifier on the satellite dish often carries d.c. power up to the low noise amplifier, radio frequency signals down to the receiver, and in some cases even digital control signals for positioning the aerial.

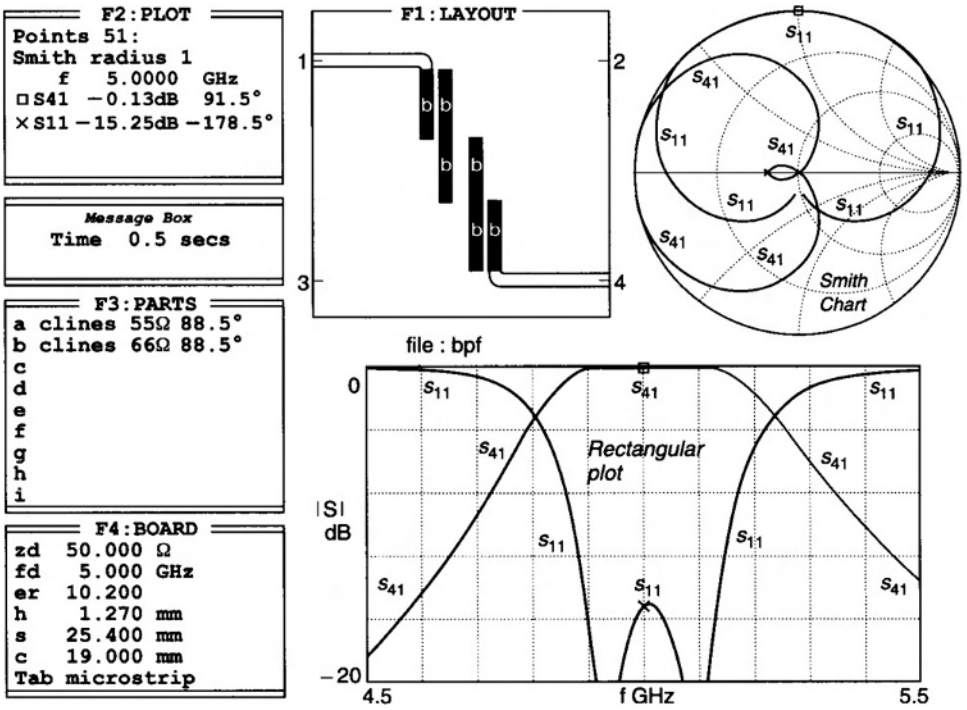


Fig. 2.2-2 Bandpass filter design using PUFF.

frequency attenuation. Frequency attenuation is known as **dispersion.**

There is also a speed or **propagation velocity** with which your voice will travel down the tunnel. This velocity is dependent on the material (air mixture of gases), its density and temperature within the tunnel. With sound waves, this velocity is about 331 metres per second.

If the tunnel is infinitely long, your voice will propagate along the tunnel until it is totally attenuated **or absorbed.** If the tunnel is not infinitely long, your voice will be **reflected** when it reaches the end wall of the tunnel and it will return to you as an **echo** or **reflected wave**. The ratio reflected wave/incident wave is called the **reflection coefficient.** You can prevent this reflection if it were physically possible to put some good sound absorption material at the end of the tunnel which absorbs all the incident sound. In other words, you would be creating a **matching termination** or **matched load** impedance (Z_L) which **matches** the **propagation characteristics** of an infinitely long tunnel in a tunnel of finite length. The ratio of the **received sound** relative to the **incident sound** is known as the **transmission coefficient.**

A signal travelling from a point A to another point B takes time to reach point B. This **time delay** is known as **propagation time delay** for the signal to travel from point A to point B. In fact, any signal travelling over any distance undergoes a propagation time delay.

Time propagation delay can be specified in three main ways: (i) **seconds,** (ii) **periodic time** (T) and (iii) **phase delay.** The first way is obvious, one merely has to note the time in seconds which it has taken for a signal to travel a given distance. Periodic time (T) is an interval of time; it is equal to [1/(frequency in Hz)] seconds. For example, if a 1000 Hz sinusoid requires four periodic times ($4T$) to travel a certain distance, then the time delay is $4 \times (1/1000)$ seconds or 4 milliseconds. Phase delay can be used to measure time because there are 2π radians in a period time (T). For the example of a 1000 Hz signal, a phase delay of $(4 \times 2\pi)$ radians is equivalent to four periodic times (T) or 4 milliseconds. Phase delay per unit length is usually represented by the symbol (β). It is measured in radians per metre.

Hence if we were to sum up propagation properties, there would be at least three properties which are obvious:

- **attenuation** of the signal as it travels along the line;
- the time or **phase delay** as the signal travels along the line;
- **dispersion** which is the different attenuation experienced by different frequencies as it travels along a line.

Finally, if you walked along a tunnel which produces echoes, while a friend whistled at a constant amplitude and pitch, you would notice the reflected sound **interfering** with the incident sound. In some places, the

whistle will sound louder (addition of incident and reflected signal); in other places the whistle will sound weaker (subtraction of incident and reflected signal). Provided your friend maintains the whistle at a constant amplitude and pitch, you will find that louder and weaker sounds always occur at the same locations in the tunnel. In other words, the pattern of louder and weaker sounds remains stationary and appears to be standing still. This change in sound intensity levels therefore produces a **standing wave pattern** along the length of the tunnel. The ratio of the maximum to minimum sound is known as the **standing wave ratio** (SWR). It will be shown later that the measurement of standing wave patterns is a very useful technique for describing the properties of transmission line systems.

In the above discussions, you have used knowledge gained from the university of life to understand the definitions of many transmission line terms. These definitions are not trivial because you will soon see that many of the above principles and terms also relate to electrical transmission lines. In fact, if you can spare the time, re-read the above paragraphs again just to ensure that you are fully cognisant of the terms shown in bold print.

2.2.3 Types of electrical transmission lines

Many of the terms introduced in the last section also apply to electrical transmission lines. However, you should be aware of the great difference in the velocity of sound waves (331 $m\,s^{-1}$) and the velocity of electrical waves (3×10^8 $m\,s^{-1}$ in air). There is also a great difference in frequency because audible sound waves are usually less than 20 kHz whereas radio frequencies are often in tens of GHz. For example, satellite broadcasting uses frequencies of about 10–12 GHz. Since wavelength = velocity/frequency, it follows that there will be a difference in wavelength and that in turn will affect the physical size of transmission lines. For example, the dimensions of a typical waveguide (Figure 2.2-3(a)) for use at frequencies between 10 GHz and 15 GHz are A = 19 mm, B = 9.5 mm, C = 21.6 mm, D = 12.1 mm.

There are many types of transmission lines.[2] These range from the two wire lines which you find in your home for table lamps, and three wire lines used for your electric kettle. Although these cables work efficiently at power frequencies (50–60 Hz), they become very inefficient at high frequencies because their inherent construction blocks high frequency signals and encourages radiation of energy.

2.2.3.1 Waveguides and coplanar waveguides

Other methods must be used and one method that comes readily to mind is the tunnel or waveguide described in Section 2.2.2.2. This waveguide is shown in Figure 2.2-3(a). It works efficiently as a high frequency transmission line because of its low attenuation and radiation losses but it is expensive because of its metallic construction (usually copper). It is also relatively heavy and lacks flexibility in use because special arrangements must be used to bend a transmission path. One variant of the waveguide is known as the coplanar waveguide (Figure 2.2-3(b)).

2.2.3.2 Coaxial and strip lines

Another way of carrying high frequency signals is to use a coaxial transmission line similar to the one that connects your TV set to its antenna. The coaxial line is shown in Figure 2.2-4(a). This is merely a two wire line but the outer conductor forms a circular shield around the inner conductor to prevent radiation.

One variation of the coaxial line appears as the strip line (Figure 2.2-4(b)). The strip line is similar to

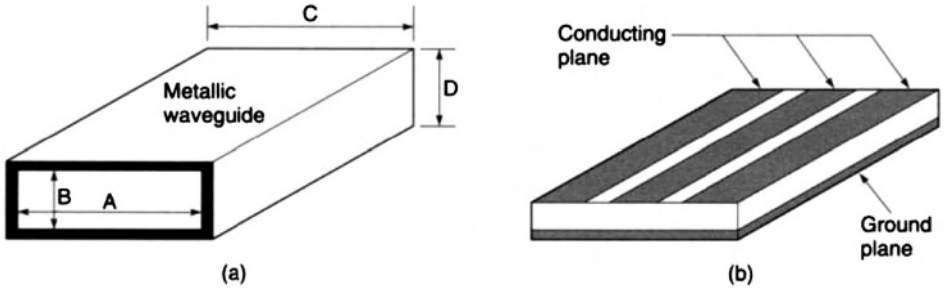


Fig. 2.2-3 (a) Metallic waveguide; (b) coplanar waveguide.

[2] The term 'transmission line' is often abbreviated to 'tx lines'.

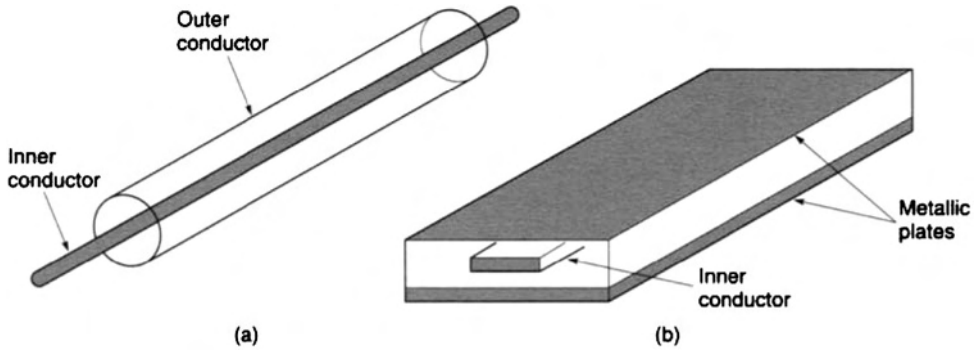


Fig. 2.2-4 (a) Coaxial cable; (b) strip line.

a 'flattened' coaxial line. It has the advantage that it can be easily constructed with integrated circuits.

2.2.3.3 Microstrip and slot lines

The microstrip line (Figure 2.2-5(a)) is a variant of the stripline with part of the 'shield' removed. The slot line (Figure 2.2-5(b)) is also a useful line for h.f. transmission.

2.2.3.4 Twin lines

In Figure 2.2-6, we show a sketch of a twin line carefully spaced by a polyethylene dielectric. This is used at relatively low frequencies. This twin cable is designed to have a characteristic impedance (Z_0) of approximately 300 Ω and it is frequently used as a very high frequency (VHF) cable or as a dipole antenna for FM radio receivers in the FM band. The parallel wire line arrangement of Figure 2.2-6 without a dielectric support can also be seen mounted on poles as overhead telephone lines, overhead power lines, and sometimes as lines connecting high power, low and medium frequency radio transmitters to their antennas.

All seven transmission lines shown in Figures 2.2-3–2.2-6 have advantages and disadvantages. For minimum loss, you would use the waveguide, the coaxial line and the strip line in integrated circuits. However, the latter two lines present difficulties in connecting external components to the inner conductor. The coplanar waveguide is better in this respect and finds favour in monolithic microwave integrated circuits (MMICs) because it allows easy series and parallel connections to external electrical components. The microstrip line is also useful for making series connections but not parallel connections because the only way through to the ground plane is either through or around the edge of the substrate. This is particularly true when a short circuit is required between the upper conductor and the ground plane; holes have to be drilled through the substrate. Microstrip also suffers from radiation losses. Nevertheless, microstrip can be made easily and conveniently and it is therefore used extensively.

2.2.3.5 Coupled lines

Coupled lines are lines which are laid alongside each other in order to permit coupling between the two lines. One example of microstrip coupled lines is shown in the F1 layout box of Figure 2.2-2 where three sets of coupled lines are used to couple energy from input port 1 to output port 4.

2.2.4 Line characteristic impedances and physical parameters

The characteristic impedance of transmission lines is calculated in two main ways:

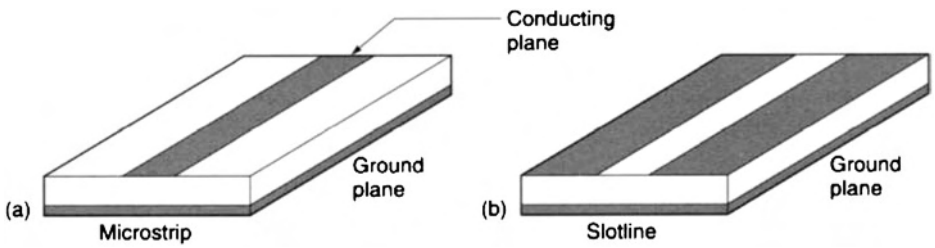


Fig. 2.2-5 (a) Microstrip line; (b) slot line.

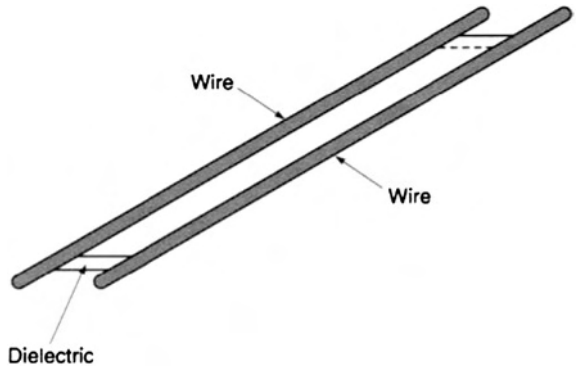


Fig. 2.2-6 Twin parallel wire VHF cable.

- from physical parameters and configuration;
- from distributed electrical parameters of the line.

Some relevant expressions for calculating the impedance of these lines from physical parameters are given in the following sections.

2.2.4.1 Coaxial line characteristic impedance (Z_0)

The expression for calculating the characteristic impedance of the coaxial transmission line shown in Figure 2.2-4(a) is

$$Z_0 = \frac{138}{\sqrt{\varepsilon}} \log_{10} \frac{D}{d} \tag{2.2.1}$$

where

d = outer diameter of the inner conductor
D = inner diameter of the outer conductor
ε = dielectric constant of the space between inner and outer conductor ($\varepsilon = 1$ for air)

Example 2.2.1

You will often find two types of flexible coaxial cables: one with a characteristic impedance Z_0 of 50 Ω which is used mainly for RF instrumentation and the other has a characteristic impedance of 75 Ω used mainly for antennas. The inner diameter of the outer conductor is the same in both cables. How would you distinguish the impedance of the two cables using only your eye?

Solution

In general, to save money, both cables are normally made with the same outer diameter. This is even more evident when the cables are terminated in a type of r.f. connector known as BNC.[3] Since these connectors have the same outer diameter, by using Equation 2.2.1 you can deduce that for $Z_0 = 75$ Ω, the inner conductor will be smaller than that of the 50 Ω cable. In practice, you will be able to recognise this distinction quite easily.

2.2.4.2 Twin parallel wire characteristic impedance (Z_0)

The expression for calculating the characteristic impedance of the type of parallel transmission line shown in Figure 2.2-6 is

$$Z_0 \approx \frac{276}{\sqrt{\varepsilon}} \log 10 \frac{2D}{d} \tag{2.2.2}$$

where

d = outer diameter of one of the identical conductors
D = distance between the centres of the two conductors
ε = relative dielectric constant = 1 for air

Example 2.2.2

The twin parallel transmission line shown in Figure 2.2-6 is separated by a distance (D) of 300 mm between the centre lines of the conductors. The diameter (d) of the identical conductors is 4 mm. What is the characteristic impedance (Z_0) of the line? Assume that the transmission line is suspended in free space, i.e. $\varepsilon = 1$.

Given: $D = 300$ mm, $d = 4$ mm, $\varepsilon = 1$.
Required: Z_0.

Solution

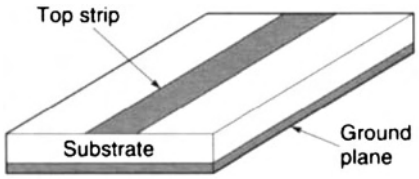
Using Equation 2.2.2

$$Z_0 \approx \frac{276}{\sqrt{1}} \log_{10} \frac{2D}{d} = \frac{276}{\sqrt{1}} \log_{10} \frac{2 \times 300}{4} \approx 600\ \Omega$$

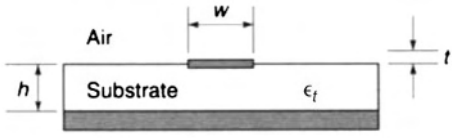
2.2.4.3 Microstrip line characteristic impedance (Z_0)

Before we start, it is best to identify some properties that are used in the calculations on microstrip. These are shown in Figure 2.2.7 where w = width of the microstrip, h = thickness of the substrate, t = thickness of the

[3] BNC is an abbreviation for 'baby N connector'. It is derived from an earlier, larger threaded connector, the Type N connector, named after Paul Neill, a Bell Laboratories engineer. BNC uses a bayonet type fixing. There is also a BNC type connector which uses a thread type fixing; it is called a TNC type connector.



(a)



(b)

Fig. 2.2-7 (a) Microstrip line; (b) end view of microstrip line.

metallisation normally assumed to approach zero in these calculations, ε_r = dielectric constant of the substrate. Note that there are two dielectric constants involved in the calculations, the relative bulk dielectric constant ε_r and the effective dielectric constant ε_e. The effective dielectric is inevitable because some of the electric field passes directly from the bottom of the strip width to the ground plane whereas some of the electric field travels via air and the substrate to the ground plate.

There are many expressions for calculating microstrip properties[4] but we will use two main methods. These are:

- an *analysis* method when we know the width/height (w/h) ratio and the bulk dielectric constant (ε_r) and want to find Z_0;
- a *synthesis* method when we know the characteristic impedance Z_0 and the bulk dielectric constant (ε_r) and want to find the w/h ratio and the effective dielectric constant (ε_e).

Analysis formulae

In the analysis case we know w/h and ε_r and want to find Z_0. The expressions which follow are mainly due to H. Wheeler's work.[5]

For *narrow* strips, i.e. $w/h < 3.3$

$$Z_0 = \frac{119.9}{\sqrt{2(\varepsilon_r+1)}}\left\{\ln\left[\frac{4h}{w}+\sqrt{16\left(\frac{h}{w}\right)^2+2}\right] - \frac{1}{2}\left(\frac{\varepsilon_r-1}{\varepsilon_r+1}\right)\left(\ln\frac{\pi}{2}+\frac{1}{\varepsilon_r}\ln\frac{4}{\pi}\right)\right\} \quad (2.2.3)$$

For *wide* strips, i.e. $w/h > 3.3$

$$Z_0 = \frac{119.9}{2\sqrt{\varepsilon_r}}\left\{\frac{w}{2h}+\frac{\ln 4}{\pi}+\frac{\ln(\varepsilon\pi^2/16)}{2\pi}\left(\frac{\varepsilon_r-1}{\varepsilon_r^2}\right) + \frac{\varepsilon_r+1}{2\pi\varepsilon_r}\left[\ln\frac{\pi\varepsilon_r}{2}+\ln\left(\frac{w}{2h}+0.94\right)\right]\right\}^{-1} \quad (2.2.4)$$

Synthesis formulae

In the synthesis case we know Z_0 and ε_r and want to find w/h and ε_e.

For *narrow* strips, i.e. $Z_0 > (44 - 2\varepsilon_r)\Omega$

$$\frac{w}{h} = \left(\frac{\exp^H}{8}-\frac{1}{4\exp^H}\right)^{-1} \quad (2.2.5)$$

where

$$H = \frac{Z_0\sqrt{2(\varepsilon_r+1)}}{119.9}+\frac{1}{2}\frac{\varepsilon_r-1}{\varepsilon_r+1}\left(\ln\frac{\pi}{2}+\frac{1}{\varepsilon_r}\ln\frac{4}{\pi}\right) \quad (2.2.6)$$

and

$$\varepsilon_e = \frac{\varepsilon_r+1}{2}+\left[1-\frac{1}{2H}\frac{\varepsilon_r-1}{\varepsilon_r+1}\left(\ln\frac{\pi}{2}+\frac{1}{\varepsilon_r}\ln\frac{4}{\pi}\right)\right]^{-2} \quad (2.2.7)$$

Note

Equation 2.2.7 was derived under a slightly different changeover value of $Z_0 > (63 - \varepsilon_r)\ \Omega$.

For *wide* strips, i.e. $Z_0 < (44 - 2\varepsilon_r)\ \Omega$

$$\frac{w}{h} = \frac{2}{\pi}[(d_e-1)-\ln(2d_e-1)]+\frac{\varepsilon_r-1}{\pi\varepsilon_r}\left[\ln(d_e-1)+0.293-\frac{0.517}{\varepsilon_r}\right] \quad (2.2.8)$$

where

$$d_e = \frac{59.95\pi^2}{Z_0\sqrt{\varepsilon_r}} \quad (2.2.9)$$

and under a slightly different value of $Z_0 > (63 - 2\varepsilon_r)\ \Omega$

$$\varepsilon_e = \frac{\varepsilon_r+1}{2}+\frac{\varepsilon_r-1}{2}\left(1+10\frac{h}{w}\right)^{-0.555} \quad (2.2.10)$$

Equations 2.2.3 to 2.2.10 are accurate up to about 2 GHz. For higher frequencies, the effect of frequency

[4] In practice these are almost always calculated using CAD/CAE programs.
[5] Wheeler, H.A. Transmission lines properties of parallel wide strips separated by a dielectric sheet, *IEEE Trans*, MTT-13 No. 3, 1965.

dependence of ε_e has to be taken into account. An expression often used to evaluate $\varepsilon_e(f)$ as frequency (f) varies is

$$\varepsilon_e(f) = \varepsilon_r - \frac{\varepsilon_r - \varepsilon_e}{1 + (h/Z_0)^{1.33}(0.43f^2 - 0.009f^3)} \quad (2.2.11)$$

where h is in millimetres, f is in gigahertz, and ε_e is the value calculated by either Equation 2.2.7 or 2.2.10.

Example 2.2.3

Two microstrip lines are printed on the same dielectric substrate. One line has a wider centre strip than the other. Which line has the lower characteristic impedance? Assume that there is no coupling between the two lines.

Solution

If you refer to Equation 2.2.3 and examine the h/w ratio, you will see that Z_0 varies as a function of h/w. Therefore, the line with the lower characteristic impedance will have a wider centre conductor.

As you can see for yourself, Equations 2.2.3 to 2.2.11 are rather complicated and should be avoided when possible. To avoid these types of calculations, we have included with this book a computer software program called PUFF. With this program, it is only necessary to decide on the characteristic impedance of the microstrip or stripline which we require and PUFF will do the rest. We will return to PUFF when we have explained the basic terms for using it.

Expressions also exist for calculating the characteristic impedance of other lines such as the strip line, coplanar waveguide, slot line, etc. These are equally complicated but details of how to calculate them have been compiled by Gupta, Garg and Chadha.[6] There is also a software program called AppCAD[7] which calculates these impedances.

Note

In the previous sections, I have produced equations which are peculiar to types of different transmission lines. From now on, and unless stated otherwise, all the equations in the sections that follow apply to all types of transmission lines.

2.2.5 Characteristic impedance (Z_0) from primary electrical parameters

A typical twin conductor type transmission line is shown in Figure 2.2-8. Each wire conductor has resistance and inductance associated with it. The resistance is associated with the material of the metal conductors, effective conductor cross-sectional area and length. The inductance is mainly dependent on length and type of material. In addition to these, there is capacitance between the two conductors. The capacitance is mainly dependent on the dielectric type, its effective permittivity, the effective cross-sectional area between conductors, the distance between the conductors and the length of the transmission line. When a voltage is applied, there is also a leakage current between the two conductors caused by the non-infinite resistance of the insulation between the two conductors. This non-infinite resistance is usually expressed in terms of a shunt resistance or parallel conductance.

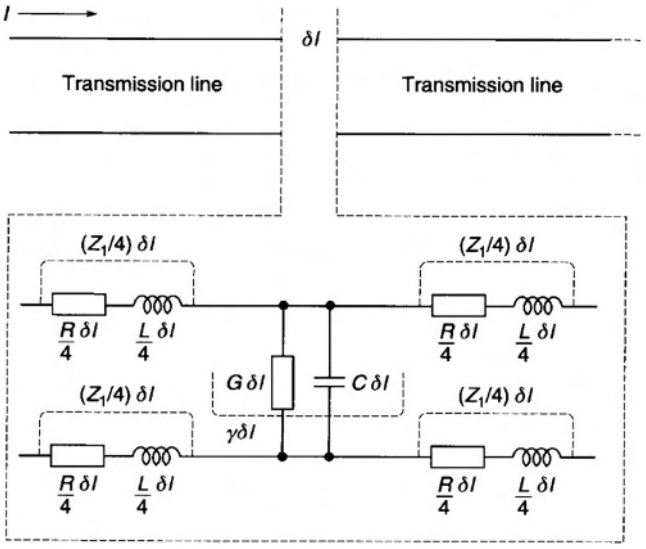


Fig. 2.2-8 Expanded view of a short section of transmission line.

Therefore, transmission lines possess inherent resistance, inductance, capacitance and conductance. It is very important to realise that these properties are distributed along the length of the line and that they are not physically lumped together. The lumped approach is only applicable when extremely short lengths of line are considered and as a practical line is made up of many short lengths of these lines, the lumped circuit equivalent of a transmission line would look more like that shown in Figure 2.2-8. This is an approximation but nevertheless it is an extremely useful one because it allows engineers to construct and simulate the properties of transmission lines.

[6] Gupta, K.C., Garg, R. and Chadha, R., *Computer-Aided Design of Microwave Circuits*, Artech House Inc, Norwood MA 02062 USA, ISBN: 0-89006-105-X.
[7] AppCAD is a proprietary software program from the Hewlett Packard Co, Page Mill Road, Palo Alto CA, USA.

2.2.5.1 Representation of primary line constants

In Figure 2.2-8, let:

R represent the resistance per metre (ohms/metre)
L represent the inductance per metre (henry/metre)
G represent the conductance per metre (siemen/metre)
C represent the capacitance per metre (farad/metre)

It follows that for a short length δl, we would obtain $R\delta l$, $L\delta l$, $G\delta l$, and $C\delta l$ respectively. Hence

$$\frac{Z_1}{4}\delta l = \left(\frac{R}{4} + \frac{j\omega L}{4}\right)\delta l \tag{2.2.12}$$

$$Z_1\delta l = (R + j\omega L)\delta l \tag{2.2.13}$$

$$Y\delta l = (G + j\omega C)\delta l \tag{2.2.14}$$

and

$$Z\delta l = \frac{1}{Y\delta l} = \frac{1}{(G + j\omega C)\delta l} \tag{2.2.15}$$

2.2.5.2 Derivation of line impedance

The input impedance Z_{in} of the short section δl when terminated by a matched line is given by[8]

$$Z_{in} = \frac{Z_1}{4}\delta l + \frac{Z\delta l\left(\frac{Z_1}{4}\delta l + Z_0 + \frac{Z_1}{4}\delta l\right)}{Z\delta l + \frac{Z_1}{4}\delta l + Z_0 + \frac{Z_1}{4}\delta l} + \frac{Z_1}{4}\delta l$$

Since the line is terminated by another line, $Z_{in} = Z_0$

$$Z_{in} = Z_0 = \frac{Z_1}{2}\delta l + \frac{Z\delta l\left(\frac{Z_1}{2}\delta l + Z_0\right)}{Z\delta l + \frac{Z_1}{2}\delta l + Z_0}$$

Cross-multiplying, we get

$$Z_0\left(Z\delta l + \frac{Z_1}{2}\delta l + Z_0\right) = \frac{Z_1}{2}\delta l\left(Z\delta l + \frac{Z_1}{2}\delta l + Z_0\right) + Z\delta l\left(\frac{Z_1}{2}\delta l + Z_0\right)$$

Simplifying

$$Z_0 Z\delta l + \frac{Z_0 Z_1}{2}\delta l + Z_0^2 = \frac{ZZ_1}{2}\delta l^2 + \frac{Z_1^2}{4}\delta l^2 + \frac{Z_0 Z_1}{2}\delta l + \frac{ZZ_1}{2}\delta l^2 + ZZ_0\delta l$$

and

$$Z_0^2 = ZZ_1\delta l^2 + \frac{Z_1^2}{4}\delta l^2 \tag{2.2.16}$$

Substituting for Z and Z_1, we get

$$Z_0^2 = \frac{(R + j\omega L)\delta l}{(G + j\omega C)\delta l} + \left(\frac{R}{4} + \frac{j\omega l}{4}\right)^2\delta l^2$$

In the limit when $\delta l \to 0$, and taking the positive square root term

$$Z_0 = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \tag{2.2.17}$$

If you examine the expression for Z_0 a bit more closely, you will see that there are two regions where Z_0 tends to be resistive and constant. The first region occurs at very low frequencies when $R \gg j\omega L$ and $G \gg j\omega C$. This results in

$$Z_0 \approx \sqrt{\frac{R}{G}} \tag{2.2.18}$$

The second region occurs at very high frequencies when $j\omega L \gg R$ and $j\omega C \gg G$. This results in

$$Z_0 \approx \sqrt{\frac{L}{C}} \tag{2.2.19}$$

The second region is also known as the frequency region where a transmission line is said to be 'lossless' because there are 'no' dissipative elements in the line.

Equation 2.2.19 is also useful because it explains why inductive loading, putting small lumped element inductors in series with lines, is used to produce a more constant impedance for the line. The frequency regions of operation described by Equations (2.2.18) and (2.2.19) are important because under these conditions, line impedance tends to remain frequency independent and a state known as 'distortionless transmission' exists. The distortionless condition is very useful for pulse waveform/digital transmissions because in these regions,

[8] The Z_0 term in the centre fraction on the right-hand side of the equation is present because the short section of line (δl) is terminated by an additional line which presents an input impedance of Z_0.

frequency dispersion and waveform distortion tend to be minimal.

These statements can also be verified by the following practical example.

Example 2.2.4

A transmission line has the following primary constants: $R = 23\ \Omega\ \text{km}^{-1}$, $G = 4\ \text{mS km}^{-1}$, $L = 125\ \mu\text{H km}^{-1}$ and $C = 48\ \text{nF km}^{-1}$. Calculate the characteristic impedance, Z_0, of the line at a frequency of (a) 100 Hz, (b) 500 Hz, (c) 15 kHz, (d) 5 MHz and (e) 10 MHz.

Given: $R = 23\ \Omega\ \text{km}^{-1}$, $G = 4\ \text{mS km}^{-1}$, $L = 125\ \mu\text{H km}^{-1}$ and $C = 48\ \text{nF km}^{-1}$.
Required: Z_0 at (a) 100 Hz, (b) 500 Hz, (c) 15 kHz, (d) 5 MHz and (e) 10 MHz.

Solution

Use Equation 2.2.17 in the calculations that follow,

(a) At 100 Hz

$$R + j\omega L = (23 + j0.08)\Omega\ \text{km}^{-1},$$
$$G + j\omega C = (4 + j0.030)\text{mS km}^{-1}$$

Hence

$$Z_0 = \sqrt{\frac{23 + j0.08}{(4 + j0.030) \times 10^{-3}}} = 75.83\Omega \angle -2.06 \times 10^{-3}\ \text{rad}$$

(b) At 500 Hz

$$R + j\omega L = (23 + j0.39)\Omega\ \text{km}^{-1},$$
$$G + j\omega C = (4 + j0.15)\text{mS km}^{-1}$$

Hence

$$Z_0 = \sqrt{\frac{23 + j0.39}{(4 + j0.15) \times 10^{-3}}} = 75.81\Omega \angle -10.30 \times 10^{-3}\ \text{rad}$$

(c) At 15 kHz

$$R + j\omega L = (23 + j11.78)\Omega\ \text{km}^{-1},$$
$$G + j\omega C = 4 + j4.52\ \text{mS km}^{-1}$$

Hence

$$Z_0 = \sqrt{\frac{23 + j11.78}{(4 + j4.52) \times 10^{-3}}} = 65.42\Omega \angle -0.19\ \text{rad}$$

(d) At 5 MHz

$$R + j\omega L = (23 + j3926.99)\Omega\ \text{km}^{-1},$$
$$G + j\omega C = (4 + j1508)\ \text{mS km}^{-1}$$

Hence

$$Z_0 = \sqrt{\frac{23 + j39.26.99}{(4 + j1508) \times 10^{-3}}} = 50.03\Omega \angle -0.00\,\text{rad}$$

(e) At 10 MHz

$$R + j\omega L = (23 + j7853.998)\Omega\ \text{km}^{-1},$$
$$G + j\omega C = (4 + j3016)\ \text{mS km}^{-1}$$

Hence

$$Z_0 = \sqrt{\frac{23 + j7853.98}{(4 + j3016) \times 10^{-3}}} = 50.03\Omega \angle -0.00\,\text{rad}$$

Conclusions from Example 2.2.4. At low frequencies, i.e. 100–500 Hz, the line impedance Z_0 tends to remain at about 75 Ω with very little phase shift over a wide frequency range. For most purposes, it is resistive and constant in this region. See cases (a) and (b). At high frequencies, i.e. 5–10 MHz, the line impedance Z_0 tends to remain constant at about 50 Ω with little phase shift over a wide frequency range. For most purposes, it is resistive and constant in this region. See cases (d) and (e). In between the above regions, the line impedance Z_0 varies with frequency and tends to be reactive. See case (c). For radio work, we tend to use transmission lines in the 'lossless' condition (Equation 2.2.19) and this helps considerably in the matching of line impedances.

2.2.6 Characteristic impedance (Z_0) by measurement

Occasions often arise when the primary constants of a line are unknown yet it is necessary to find the characteristic impedance (Z_0). In this case, Z_0 can be obtained by measuring the short- and open-circuit impedance of the line. In Figures 2.2-9 and 2.2-10 as in Figure 2.2-8 let:

R represent the resistance per metre (ohms/metre)
L represent the inductance per metre (henry/metre)

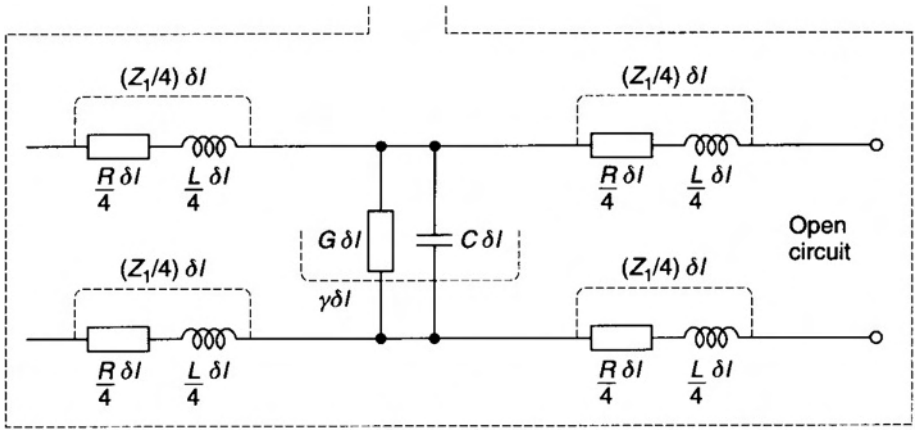


Fig. 2.2-9 Open-circuit equivalent of a short length of transmission line.

G represent the conductance per metre (siemen/metre)
C represent the capacitance per metre (farad/metre)

It follows that for a short length δl, we would obtain $R\delta l$, $L\delta l$, $G\delta l$ and $C\delta l$ respectively.

2.2.6.1 Open-circuit measurement (Z_{oc})

Hence defining $Z\delta l$ as $1/Y\delta l$ we have

$$Z_{oc} = \frac{Z_1}{4}\delta l + Z\delta l + \frac{Z_1}{4}\delta l = \frac{Z_1}{2}\delta l + Z\delta l \tag{2.2.20}$$

2.2.6.2 Short-circuit measurement (Z_{sc})

The short-circuit impedance is

$$Z_{sc} = \frac{Z_1}{4}\delta l + \frac{Z\delta l\left(\frac{Z_1}{4}\delta l + \frac{Z_1}{4}\delta l\right)}{Z\delta l + \left(\frac{Z_1}{4}\delta l + \frac{Z_1}{4}\delta l\right)} + \frac{Z_1}{4}\delta l$$

$$= \frac{Z_1}{2}\delta l + \frac{\frac{ZZ_1}{2}\delta l^2}{Z\delta l + \frac{Z_1}{2}\delta l}$$

$$= \frac{\frac{Z_1}{2}\delta l\left(Z\delta l + \frac{Z_1}{2}\delta l\right) + \frac{ZZ_1}{2}\delta l^2}{Z\delta l + \frac{Z_1}{2}\delta l} = \frac{ZZ_1\delta l^2 + \frac{Z_1^2}{4}\delta l^2}{Z\delta l + \frac{Z_1}{2}\delta l}$$

Using Equation 2.2.16 to substitute for the numerator and Equation 2.2.20 to substitute for the denominator, we have

$$Z_{sc} = \frac{Z_0^2}{Z_{oc}}$$

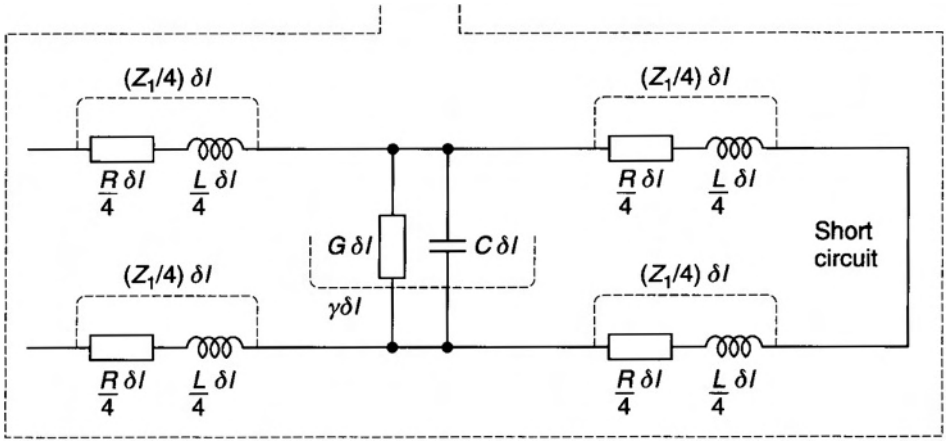


Fig. 2.2-10 Short-circuit equivalent of a short length of transmission line.

Hence

$$Z_0^2 = Z_{sc}Z_{oc}$$

or

$$Z_0 = \sqrt{Z_{sc}Z_{oc}} \qquad (2.2.21)$$

Example 2.2.5

The following measurements have been made on a line at 1.6 MHz where $Z_{oc} = 900\,\Omega \angle -30^\circ$ and $Z_{sc} = 400\,\Omega \angle -10^\circ$. What is the characteristic impedance (Z_0) of the line at 1.6 MHz?

Given: $f = 1.6$ MHz,
$Z_{oc} = 900\,\Omega \angle -30^\circ, Z_{sc} = 400\,\Omega \angle -10^\circ$.
Required: Z_0 at 1.6 MHz.

Solution

Using Equation 2.2.21

$$Z_0 = \sqrt{Z_{sc}Z_{oc}} = \sqrt{900\,\Omega \angle -30^\circ \times 400\,\Omega \angle -10^\circ}$$
$$= 600\,\Omega \angle -20^\circ$$

2.2.7 Typical commercial cable impedances

Manufacturers tend to make cables with the following characteristic impedances (Z_0). These are:

- 50 Ω – This type of cable finds favour in measurement systems and most radio instruments are matched for this impedance. It is also used extensively in amateur radio links.

  Most cable manufacturers make more than one type of 50 Ω line. For example, you can buy 50 Ω rigid lines (solid outer connector), 50 Ω low loss lines (helical and air dielectrics), 50 Ω high frequency lines for use up to 50 GHz with minimal loss. The reason for this is that different uses require different types of lines. Remember that in Equation 2.2.1 repeated here for convenience $Z_0 = 138\log_{10}(D/d)/\sqrt{\varepsilon}$ and the dimension of the variables can be changed to produce the desired impedance.
- 75 Ω – This type of cable is favoured by the television industry because it provides a close match to the impedance (73.13 Ω) of a dipole aerial. Most TV aerials are designed for this impedance and it is almost certain that the cable that joins your TV set to the external aerial will have this impedance. The comments relating to the different types of 50 Ω lines also apply to 75 Ω lines.
- 140 Ω – This type of cable is used extensively by the telephone industry. The comments relating to the different types of 50 Ω lines also apply to 140 Ω lines.
- 300 Ω – This type of cable is favoured by both the radio and television industry because it provides a close match for the impedance (292.5 Ω) of a very popular antenna (folded dipole antenna) which is used extensively for VHF-FM reception. The comments relating to the different types of 50 Ω lines also apply to 300 Ω lines.
- 600 Ω – This type of cable is used extensively by the telephone industry and many of their instruments are matched to this impedance. The comments relating to the different types of 50 Ω lines also apply to 600 Ω lines.

2.2.8 Signal propagation on transmission lines

2.2.8.1 Pulse propagation on an infinitely long or matched transmission line

We are now going to use some of the ideas introduced in the previous sections, particularly Section 2.2.2.2, to describe qualitatively the propagation of signals along an infinitely long transmission line. In this description we will only make two assumptions:

- the transmission line is perfectly uniform, that is its electrical properties are identical all along its length;
- the line extends infinitely in one direction or is perfectly terminated.

To keep the explanation simple, we will initially only consider the propagation of a single electrical pulse along the line[9] shown in Figure 2.2-11. At the beginning of the

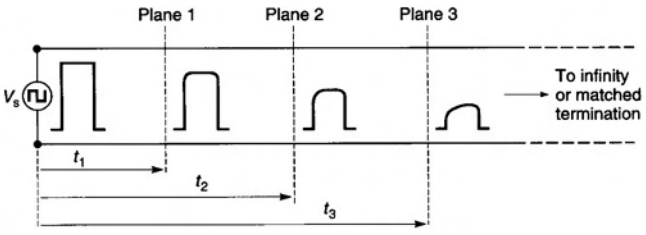


Fig. 2.2-11 Pulse propagation in a transmission line.

[9] The behaviour of a pulse travelling along an infinitely long transmission line is very similar to the example you were given in Section 2.2.2.2 concerning sound travelling down an infinitely long tunnel except that this time instead of voice sounds, consider the sound to originate from a single drum beat or pulse. You will no doubt remember from earlier work that a pulse is a waveform which is made up from a fundamental sinusoid and its harmonics combined together in a precise amplitude, phase and time relationship.

line (top left hand corner) a voltage source (V_s) produces the single pulse shown in Figure 2.2-11. The waveforms shown at various planes (plane 1, plane 2, plane 3) on the line illustrate three of the main properties of signal propagation along a transmission line:

- **propagation delay** – the pulse appears at each successive point on the line later than at the preceding point;
- **attenuation** – the peak value of the pulse is attenuated progressively;
- **waveform distortion and frequency dispersion** – its shape differs from its original shape at successive points.

2.2.8.2 Propagation delay

The pulse appears later and later at successive points on the line because it takes time to travel over any distance, i.e. there is a propagation delay. As the line is uniform throughout its length, the amount of delay at any point is proportional to distance between that point and the source of the pulse. These time delays are shown as t_1, t_2, and t_3 in Figure 2.2-11. Another way of describing this is to say that the pulse propagates along the line with a uniform velocity.

2.2.8.3 Attenuation

The amplitude of the pulse is attenuated as it propagates down the line because of resistive losses in the wires. The amount of attenuation per unit length is uniform throughout the line because the line cross-section is uniform throughout the line length. Uniform attenuation means that the fractional reduction in pulse amplitude is the same on any line section of a given length. This is more easily understood by referring to Figure 2.2-11, where the pulse amplitude at plane 1 has been reduced by a factor of 0.8. At plane 2, which is twice as far from the source as plane 1, the pulse height has been reduced by a further factor of 0.8, i.e. a total of 0.8^2 or 0.64 of its original amplitude. At plane 3, which is three times as far from the source as plane 1, the reduction is 0.8^3 or 0.512 of the original amplitude.

More generally, at a distance equal to l times the distance from the source to plane 1, the height is reduced by $(0.8)^l$. Because l is the exponent in this expression this type of amplitude variation is called exponential. It can also be expressed in the form $(e^{\alpha})^l$ or $e^{\alpha l}$, where e^{α} represents the loss per unit length and is 0.8 in this particular example. In fact α is the natural logarithm of the amplitude reduction per unit length. Its unit is called the neper and loss (dB) = 8.686 nepers.[10]

Example 2.2.6

A transmission has a loss of two nepers per kilometre. What is the loss in dB for a length of 10 kilometres?

Given: Attenuation constant (α) = 2 nepers per km.
Required: Loss in dB for a length of 10 km.

Solution

If 1 km represents a loss of 2 nepers, then 10 km = 10 × 2 = 20 nepers. Therefore

$$\text{loss} = 8.686 \times 20 = 173.72\,\text{dB}$$

2.2.9 Waveform distortion and frequency dispersion

2.2.9.1 Amplitude distortion

The waveform of the pulse in Figure 2.2-11 alters as it travels along the line. This shape alteration is caused by the line constants (inductance, capacitance, resistance and conductance of the line) affecting each sinusoidal component of the waveform in a different manner. The high frequency components, which predominate on the edges of the pulse waveform, suffer greater attenuation because of increased reactive effects; the lower frequency components, which predominate on the flat portion of the waveform, suffer less attenuation. The variation of attenuation with frequency is described by the frequency response of the line.

2.2.9.2 Frequency distortion

In addition to attenuation, there are also time constants associated with the line components (inductance, capacitance, resistance and conductance). These cause high frequency components to travel at a different velocity from low frequency components. The variation of velocity with frequency is called the frequency dispersion of the line.

2.2.9.3 Phase and group velocities

As a pulse consists of sinusoidal components of different frequencies, each component will therefore be altered differently. Distinction must be made between the velocities of the sinusoidal components which are called

[10] This is because dB = 20 Log (Ratio) = 20 Log (e^{α}) = 20 × α × Log (e) = 20 × α × 0.4343 = 8.686α.

phase velocities, $\boldsymbol{u}_{\mathbf{p}}$. The phase velocity ($\beta$) is defined as the change in radians over a wavelength and since there is a phase change of 2π radians in every wavelength, it follows that $\beta = 2\pi$ radians/wavelength (λ) or

$$\beta = \frac{2\pi}{\lambda} \tag{2.2.22}$$

The velocity of the complete waveform is called the group velocity, $\boldsymbol{u}_{\mathrm{g}}$. The apparent velocity of the pulse in Figure 2.2-11 is called its group velocity.

It is important to realise that if the line velocity and line attenuation of all the component sinusoids which make up a pulse waveform are not identical then deterioration in pulse waveform shape will occur. Pulse distortion is particularly critical in high speed data transmission where a series of distorted pulses can easily merge into one another and cause pulse detection errors.

If distortion occurs and if it is desired to know how and why a particular waveform has changed its shape, it will be necessary to examine the propagation of the constituent sinusoids of the waveform itself and to instigate methods, such as frequency and phase equalisation, to ensure minimal waveform change during signal propagation through the line.

2.2.10 Transmission lines of finite length

2.2.10.1 Introduction

In Section 2.2.8.1, we discussed waveforms travelling down infinitely long lines. In practice, infinitely long lines do not exist but finite lines can be made to behave like infinitely long lines if they are terminated with the characteristic impedance of the line.[11]

2.2.10.2 Matched and unmatched lines

A transmission line which is terminated by its own characteristic impedance, Z_0, is said to be **matched** or **properly terminated.** A line which is terminated in any impedance other than Z_0 is said to be **unmatched** or **improperly terminated.** To prevent reflections it is usual for a transmission line to be properly terminated and so it is a common condition for a transmission line to behave electrically as though it was of infinite length.

If a transmission line is to be used for signals with a wide range of frequency components, it may be difficult to terminate it properly. In general, the characteristic impedance of a transmission line will vary with frequency and if the matching load fails to match the line at all frequencies, then the line will not be properly terminated and reflections will occur.[12]

In practice, it is usual to properly terminate both ends of a transmission line, i.e. both at the sending end and the receiving end; otherwise any signal reflected from the receiving end and travelling back towards the sending end will be re-reflected again down the line to cause further reflections. The sending end can be properly terminated either by using a source generator with an impedance equal to the characteristic impedance of the line or by using a matching network to make a source generator present a matched impedance to the transmission line.

2.2.11 Reflection transmission coefficients and VSWR

2.2.11.1 Introduction

Reflection coefficients are based on concepts introduced in your childhood. Consider the case when you throw a ball at a vertical stone wall. The ball with its incident power will travel towards the wall, hit the wall which will absorb some of its incident power and then the remaining power (reflected power) will cause the ball to bounce back.

The ratio (reflected power)/(incident power) is called the reflection coefficient. The reflection coefficient is frequently represented by the Greek letter gamma (Γ). In mathematical terms, we have

$$\Gamma = \frac{\text{reflected power}}{\text{incident power}}$$

This simple equation is very useful for the following reasons.

- Its value is independent of incident power because if you double incident power, reflected power will also double.[13] If you like, you can say that Γ is normalised to its incident power.
- It gives you a measure of the hardness **(impedance)** of the wall to incident power. For example if the wall is made of stone, it is likely that very little

[11] This argument is similar to the case mentioned in Section 2.2.2.2 where it was shown that if our finite tunnel was terminated with material with the same properties as an infinitely long tunnel which absorbed all the incident energy then it would also behave like an infinitely long tunnel.
[12] You see this reflection effect as multiple images on your television screen when the TV input signal is not properly terminated by the TV system. TV engineers call this effect 'ghosting'.
[13] This of course assumes that the hardness of your wall is independent of the incident power impinging on it.

incident power will be absorbed and most of the incident power will be returned to you as reflected power. You will get a high reflection coefficient ($\Gamma \rightarrow 1$). If the wall is made of wood, it is likely that the wood would bend a bit (less resistance), absorb more incident energy and return a weaker bounce. You will get a lower reflection coefficient ($\Gamma < 1$). Similarly if the wall was made of straw, it is more than likely that most of the incident energy would be absorbed and there would be little rebounce or reflected energy ($\Gamma \rightarrow 0$). Finally if the wall was made of air, the ball will simply go through the air wall and carry all its incident power with it ($\Gamma = 0$). There will be no reflected energy because the incident energy would simply be expended in carrying the ball further. Note in this case that the transmission medium resistance is air, and it is the same as the air wall resistance which is the load and we simply say that the load is matched to the transmission medium.

- By measuring the angle of the re-bounce relative to the incident direction, it is possible to tell whether the wall is vertical and facing the thrower or whether it is at an angle (phase) to the face-on position. Hence we can determine the direction of the wall.
- The path through which the ball travels is called the **transmission path.**
- Last but not least, you need not even physically touch the wall to find out some of its characteristics. In other words, measurement is *indirect*. This is useful in the measurement of transistors where the elements cannot be directly touched. It is also very useful when you want to measure the impedance of an aerial on top of a high transmitting tower when your measuring equipment is at ground level. The justification for this statement will be proved in Section 2.2.13.

2.2.11.2 Voltage reflection coefficient[14] (Γ_v) in transmission lines

The same principles described above can also be applied to electrical energy. This is best explained by Figure 2.2.12 where we have a signal generator with a source impedance, Z_s, sending electrical waves through a transmission line whose impedance is Z_0, into a load impedance, Z_L.

If the load impedance (Z_L) is exactly equal to Z_0, the incident wave is totally absorbed in the load and there is no reflected wave. If Z_L differs from Z_0, some of the incident wave is not absorbed in the load and is reflected

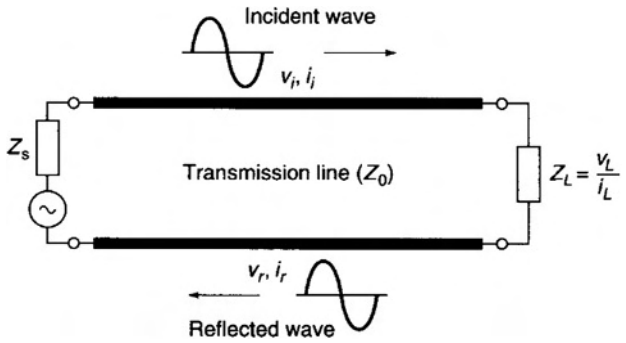


Fig. 2.2-12 Incident and reflected waves on a transmission line.

back towards the source If the source impedance (Z_s) is equal to Z_0, the reflected wave from the load will be absorbed in the source and no further reflections will occur. If Z_s is not equal to Z_0, a portion of the reflected wave from the load is re-reflected from the source back toward the load and the entire process repeats itself until all the energy is dissipated. The degree of mismatch between Z_0 and Z_L or Z_s determines the amount of the incident wave that is reflected.

By definition

$$\text{voltage reflection coefficient} = \frac{v_{\text{reflected}}}{v_{\text{incident}}} = \Gamma_v \angle \theta \tag{2.2.23}$$

Also

$$\text{current reflection coefficient} = \frac{i_{\text{reflected}}}{i_{\text{incident}}} = \Gamma_i \angle \theta \tag{2.2.24}$$

From inspection of the circuit of Figure 2.2-12

$$Z_0 = \frac{v_i}{i_i} \tag{2.2.25}$$

and

$$Z_0 = \frac{v_r}{i_r} \tag{2.2.26}$$

The minus sign in Equation 2.2.27 occurs because we use the mathematical convention that current flows to the right are positive; therefore, current flows to the left are negative.

$$Z_L = \frac{v_L}{I_L} = \frac{v_i + v_r}{i_i - i_r} = \frac{v_i + v_r}{v_i/Z_0 - v_r/Z_0} = Z_0 \frac{v_i(1+\Gamma_v)}{v_i(1-\Gamma_v)} \tag{2.2.27}$$

[14] Some authors use different symbols for voltage reflection coefficient. Some use Γ_v, while others use ρ_v. In this book, where possible, we will use Γ_v for components and ρ_v for systems.

Sorting out terms in respect of Γ_v

$$Z_L = Z_0\frac{(1+\Gamma_v)}{(1-\Gamma_v)}$$

or

$$\Gamma_v = \frac{(Z_L - Z_0)}{(Z_L + Z_0)} \tag{2.2.28}$$

Returning to Equation 2.2.24 and recalling Equation 2.2.23

$$\Gamma_i = \frac{i_r}{i_i} = \frac{-v_r/Z_0}{v_i/Z_0} = -\Gamma_v \tag{2.2.29}$$

As the match between the characteristic impedance of the transmission line Z_0 and the terminating impedance Z_L improves, the reflected wave becomes smaller. Therefore, using Equation 2.2.28, the reflection coefficient decreases. When a perfect match exists, there is no reflected wave and the reflection coefficient is zero. If the load Z_L on the other hand is an open or short circuit, none of the incident power can be absorbed in the load and all of it will be reflected back toward the source. In this case, the reflection coefficient is equal to 1, or a *perfect* mismatch. Thus the *normal* range of values for the *magnitude* of the reflection coefficient is between zero and unity.

Example 2.2.7

Calculate the voltage reflection coefficient for the case where $Z_L = (80 - j10)\Omega$ and $Z_0 = 50\ \Omega$.

Given: $Z_L = (80 - j10)$, $Z_0 = 50\,\Omega$
Required: Γ_v

Solution

Using Equation 2.2.28

$$\Gamma_v = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{80 - j10 - 50}{80 - j10 + 50} = \frac{30 - j10}{130 - j10}$$
$$= \frac{31.62\angle -18.43^\circ}{130.38\angle -4.40^\circ} = 0.24\ \angle -14.03^\circ$$

Example 2.2.8

Calculate the voltage reflection coefficients at the terminating end of a transmission line with a characteristic impedance of 50 Ω when it is terminated by (a) a 50 Ω termination, (b) an open-circuit termination, (c) a short-circuit termination and (d) a 75 Ω termination.

Given: $Z_0 = 50\ \Omega$, Z_L = (a) 50 Ω, (b) open-circuit = ∞, (c) short-circuit = 0 Ω, (d) = 75 Ω.

Required: Γ_v for (a), (b), (c), (d).

Solution

Use Equation 2.2.28.

(a) with $Z_L = 50\ \Omega$

$$\Gamma_v = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{50 - 50}{50 + 50} = 0\,\underline{/0^\circ}$$

(b) with Z_L = open circuit = $\infty\ \Omega$

$$\Gamma_v = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{\infty - 50}{\infty + 50} = 1\,\underline{/0^\circ}$$

(c) with Z_L = short circuit = 0 Ω

$$\Gamma_v = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{0 - 50}{0 + 50} = -1\,\underline{/0^\circ} \text{ or } 1/\underline{180^\circ}$$

(d) with $Z_L = 75\ \Omega$

$$\Gamma_v = \frac{Z_L - Z_0}{Z_L + Z_0} = \frac{75 - 50}{75 + 50} = 0.2\,\underline{/0^\circ}$$

Example 2.2.8 is instructive because it shows the following.

- If you want to transfer an incident voltage wave with no reflections then the terminating load (Z_L) must match the characteristic impedance (Z_0) exactly. See case (a). This is the desired condition for efficient transfer of power through a transmission line.
- Maximum in-phase voltage reflection occurs with an open circuit and maximum antiphase voltage reflection occurs with a short circuit. See cases (b) and (c). This is because there is no voltage across a short circuit and therefore the reflected wave must cancel the incident wave.
- Intermediate values of terminating impedances produce intermediate values of reflection coefficients. See case (d).

2.2.11.3 Return loss

Incident power (P_{inc}) and reflected power (P_{ref}) can be related by using the magnitude of the voltage reflection coefficient (Γ). Since $\Gamma = v_{ref}/v_{inc}$, it follows that

$$\frac{P_{ref}}{P_{inc}} = \frac{v_{ref}^2/R_{load}}{v_{inc}^2/R_{load}} = \Gamma^2 \tag{2.2.30}$$

The return loss gives the amount of power reflected from a load and is calculated from

$$\text{return loss (dB)} = -10\log\Gamma^2 = -20\log\Gamma \tag{2.2.31}$$

2.2.11.4 Mismatched loss

The amount of power transmitted to the load (P_L) is determined from

$$P_L = P_{inc} - P_{ref} = P_{inc}(1-\Gamma^2) \tag{2.2.32}$$

The fraction of the incident power not reaching the load because of mismatches and reflections is

$$\frac{P_{load}}{P_{incident}} = \frac{P_L}{P_{inc}} = 1-\Gamma^2 \tag{2.2.33}$$

Hence the mismatch loss (or reflection loss) is calculated from

$$\text{ML (dB)} = -10\log(1-\Gamma^2) \tag{2.2.34}$$

2.2.11.5 Transmission coefficient

The transmission coefficient (τ_v) is defined as the ratio of the load voltage (v_L) to the incident voltage (v_{inc}) but $v_L = v_{inc} + v_{ref}$. Hence

$$\tau_v = \frac{v_L}{v_{inc}} = \frac{v_{inc}+v_{ref}}{v_{inc}} = 1+\Gamma_v \tag{2.2.35}$$

If we now use Equation 2.2.28 to substitute for Γ_v, we obtain

$$\tau_v = 1+\Gamma_v = 1+\frac{Z_L-Z_0}{Z_L+Z_0} = \frac{2Z_L}{Z_L+Z_0} \tag{2.2.36}$$

Sometimes Equation 2.2.36 is normalised to Z_0 and when Z_L/Z_0 is defined as z, we obtain

$$\tau_v = \frac{2z}{z+1} \tag{2.2.36a}$$

Equation 2.2.36a is the form you frequently find in some articles.

2.2.11.6 Voltage standing wave ratio (VSWR)

Cases often arise when the terminating impedance for a transmission line is not strictly within the control of the designer. Consider a typical case where a transmitter designed for operating into a 50 Ω transmission line is made to feed an antenna with a nominal impedance of 50 Ω. In the ideal world, apart from a little loss in the transmission line, all the energy produced by the transmitter will be passed on to the antenna. In the practical world, an exact antenna match to the transmission line is seldom achieved and most antenna manufacturers are honest enough to admit the discrepancy and they use a term called the **voltage standing wave ratio** (VSWR)[15] to indicate the degree of mismatch.

VSWR is useful because

- it is relatively easy to measure – it is based on modulus values rather than phasor quantities which enables simple diode detectors to be used for measurement purposes;
- it indicates the degree of mismatch in a termination;
- it is related to the modulus of the reflection coefficient (shown later).

VSWR is defined as

$$\text{VSWR} = \frac{|V_{max}|}{|V_{min}|} = \frac{|V_{inc}|+|V_{ref}|}{|V_{inc}|-|V_{ref}|} \tag{2.2.37}$$

A VSWR of |1| represents the best possible match.[16] Any VSWR greater than |1| indicates a mismatch and a large VSWR indicates a greater mismatch than a smaller VSWR. Typical figures of VSWRs for good practical terminations range from 1.02 to 1.1.

[15] This term is based on the Standing Wave Pattern principle which was introduced in Section 2.2.2.2 where you walked along a tunnel which produced echoes while your friend whistled at a constant amplitude and pitch. In the tunnel case, the loudest (maximum intensity) sound occurred where the incident and reflected wave added, while the weakest sound (minimum intensity) occurred where the incident and reflected sound opposed each other.

[16] In a properly terminated line, there are no reflections. $V_{ref} = 0$ and substituting this value into Equation 2.2.37 gives

$$\text{VSWR} = \frac{|V_{inc}|+|0|}{|V_{inc}|-|0|} = |1|$$

Example 2.2.9

In Figure 2.2-12, the incident voltage measured along the transmission line is 100 V and the reflected voltage measured on the same line is 10 V. What is its VSWR?

Solution

Using Equation 2.2.37

$$\text{VSWR} = \frac{|V_{\text{inc}}| + |V_{\text{ref}}|}{|V_{\text{inc}}| - |V_{\text{ref}}|} = \frac{|100| + |10|}{|100| - |10|} = 1.22$$

2.2.11.7 VSWR and reflection coefficient (Γ_v)

VSWR is related to the voltage reflection coefficient by

$$\text{VSWR} = \frac{|V_{\text{inc}}| + |V_{\text{ref}}|}{|V_{\text{inc}}| - |V_{\text{ref}}|} = \frac{1 + \left|\frac{V_{\text{ref}}}{V_{\text{inc}}}\right|}{1 - \left|\frac{V_{\text{ref}}}{V_{\text{inc}}}\right|} = \frac{1 + |\Gamma_v|}{1 - |\Gamma_v|} \tag{2.2.38}$$

or

$$|\Gamma_v| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} \tag{2.2.38a}$$

Example 2.2.10

What is the VSWR of a transmission system if its reflection coefficient $|\Gamma_v|$ is 0.1?

Given: $|\Gamma_v| = 0.1$
Required: VSWR

Solution

Using Equation 2.2.38

$$\text{VSWR} = \frac{1 + |\Gamma_v|}{1 - |\Gamma_v|} = \frac{1 + |0.1|}{1 - |0.1|} = 1.22$$

Perfect match occurs when VSWR = |1| This is the optimum condition and examination of Equation 2.2.38 shows that this occurs when $|\Gamma_v| = 0$. With this condition, there is no reflection, optimum power is transferred to the load and there are no standing wave patterns on the line to cause excessive insulation breakdown or electrical discharges to surrounding conductors and there are no 'hot spots' or excessive currents on sections of the conductors.

Example 2.2.11

A manufacturer quotes a maximum VSWR of 1.07 for a resistive load when it is used to terminate a 50 Ω transmission line. Calculate the reflected power as a percentage of the incident power.

Given: VSWR = 1.07, $Z_0 = 50\ \Omega$
Required: Reflected power as a percentage of incident power

Solution

Using Equation 2.2.38a

$$|\Gamma_v| = \frac{\text{VSWR} - 1}{\text{VSWR} + 1} = 0.034$$

Since power is proportional to V^2

$$\begin{aligned} P_{\text{ref}} &= (0.034)^2 \times P_{\text{inc}} \\ &= 0.001 \times P_{\text{inc}} \\ &= 0.1\% \text{ of } P_{\text{inc}} \end{aligned}$$

From the answer to Example 2.2.11, you should now realise that:

- a load with an SWR of 1.07 is a good terminating load;
- there are hardly any reflections when a transmission line is terminated with such a load;
- the transmission line is likely to behave like an infinitely long line.

2.2.11.8 Summary of Section 2.2.11

If a transmission line is not properly terminated, reflections will occur in a line. These reflections can aid or oppose the incident wave. In high voltage lines, it is possible for the aiding voltages to cause line insulation breakdown. In high current lines, it is possible at high current points for aiding currents to overheat or even destroy the metallic conductors.

The voltage reflection coefficient (Γ_v) can be calculated by

$$\Gamma_v = \frac{\text{reflected voltage wave}}{\text{incident voltage wave}} = \frac{Z_L - Z_0}{Z_L + Z_0}$$

Manufacturers tend to use VSWRs when quoting the impedances associated with their equipment. A VSWR of |1| is the optimum condition and indicates that a perfect match is possible and that there will be no reflections when the equipment is matched to a perfect transmission line. VSWR can be calculated from the reflection coefficients by

$$\text{VSWR} = \left|\frac{V_{\max}}{V_{\min}}\right| = \frac{|V_{\text{inc}}| + |V_{\text{ref}}|}{|V_{\text{inc}}| - |V_{\text{ref}}|} = \frac{1 + |\Gamma_v|}{1 - |\Gamma_v|}$$

The return loss is a way of specifying the power reflected from a load and is equal to $-10 \log \Gamma^2$. The mismatch loss

or reflection loss specifies the fraction of incident power not reaching the load and is equal to $-10 \log (1 - \Gamma^2)$.

2.2.12 Propagation constant (γ) of transmission lines

2.2.12.1 Introduction

In Section 2.2.8, we saw that signals on transmission lines suffer attenuation, phase or time delay, and often frequency distortion. In this section, we will show the relationships between these properties and the primary constants (R, G, L and C) of a transmission line.

2.2.12.2 The propagation constant (γ) in terms of the primary constants

To find the **propagation constant** (γ) we start with the same equivalent circuit (Figure 2.2-8) used for the derivation of Z_0. It is re-drawn in Figure 2.2-13 with the voltage and current phasors indicated.

The propagation constant, as defined, relates V_2 and V_1 by

$$\frac{V_2}{V_1} = \mathrm{e}^{-\gamma \delta l} \tag{2.2.39}$$

where δl is still the short length of line referred to in Figure 2.2-8. It is easier to find γ using the current phasors rather than the voltage phasors; so, using $I_1 = V_1/Z_0$ and $I_2 = V_2/Z_0$

$$\frac{I_2}{I_1} = \mathrm{e}^{-\gamma \delta l} \tag{2.2.40}$$

or alternatively

$$\frac{I_1}{I_2} = \mathrm{e}^{\gamma \delta l} \tag{2.2.40a}$$

The current I_1 splits into two parts: I_2 and a part going through Z_2. By the current divider rule, the split is

$$I_2 = \frac{Z_2}{Z_2 + Z_1/2 + Z_0} I_1$$

giving

$$\frac{I_1}{I_2} = 1 + \frac{Z_1}{2Z_2} + \frac{Z_0}{Z_2}$$

Substituting the definitions for Z_1 and Z_2 and the formula for Z_0 derived above gives

$$\begin{aligned}\frac{I_1}{I_2} &= 1 + \frac{1}{2}(R + \mathrm{j}\omega L)(G + \mathrm{j}\omega L)(\delta l)^2 \\ &\quad + \sqrt{(R + \mathrm{j}\omega L)(G + \mathrm{j}\omega L)\delta l} \\ &= 1 + \sqrt{(R + \mathrm{j}\omega L)(G + \mathrm{j}\omega L)\delta l} \\ &\quad + \frac{1}{2}(R + \mathrm{j}\omega L)(G + \mathrm{j}\omega L)(\delta l)^2\end{aligned}$$

Also $I_1/I_2 = \mathrm{e}^{\gamma \delta l}$. To use these two expressions for I_1/I_2 to find γ we must first expand $\mathrm{e}^{\gamma \delta l}$ into a Taylor series. Since

$$\mathrm{e}^x = 1 + x + \frac{x^2}{2} + \cdots$$

we can write $\mathrm{e}^{\gamma \delta \mathrm{l}}$ as

$$\mathrm{e}^{\gamma \delta l} = 1 + \gamma \delta l + \frac{\gamma^2 (\delta l)^2}{2} + \cdots$$

Equating the two expressions for I_1/I_2 gives

$$\begin{aligned}1 + \gamma \delta l + \gamma^2 (\delta l)^2/2 &= 1 + \sqrt{(R + \mathrm{j}\omega L)(G + \mathrm{j}\omega L)\delta l} \\ &\quad + \frac{1}{2}(R + \mathrm{j}\omega L)(G + \mathrm{j}\omega L)(\delta l)^2\end{aligned}$$

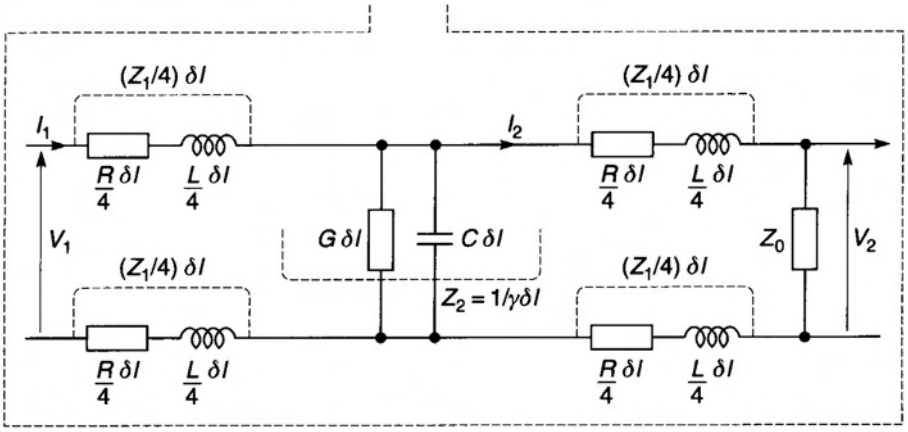


Fig. 2.2-13 Equivalent circuit of a very short length of line.

Subtracting 1 from each side and dividing by δl gives

$$\gamma\delta l + \gamma^2(\delta l)^2/2 = \sqrt{(R + j\omega L)(G + j\omega L)\delta l} + \frac{1}{2}(R + j\omega L)(G + j\omega L)(\delta l)^2$$

and as δl approaches zero

$$\gamma = \sqrt{(R + j\omega L)(G + j\omega L)} \quad (2.2.41)$$

Since γ is complex consisting of a real term α for amplitude and β for phase, we can also write

$$\gamma = \alpha + j\beta = \sqrt{(R + j\omega L)(G + j\omega L)} \quad (2.2.42)$$

If the expression for γ (Equation 2.2.42) is examined more closely, it can be seen that there are two regions where γ tends to be resistive and constant. The first region occurs at very low frequencies when $R \gg j\omega L$ and $G \gg j\omega C$. This results in

$$\gamma \approx \sqrt{(R)(G)} \quad (2.2.43)$$

In this region γ is a real number which does not depend on ω. Since the real part of γ is α, the attenuation index, there is no amplitude distortion in the very low frequency range. The second region occurs at very high frequencies when $j\omega L \gg R$ and $j\omega C \gg G$. This results in

$$\gamma \approx j\omega\sqrt{(L)(C)} \quad (2.2.44)$$

In this region γ is purely imaginary and is proportional to ω. Since the imaginary part of γ is β, the phase index, it means that there is no dispersion (because β is proportional to ω) in the high frequency range. The region is very useful for pulse waveform/digital transmissions because in it frequency dispersion and waveform distortion tend to be minimal.

Equation 2.2.44 is also useful because it explains why inductive loading, putting small lumped element inductors in series with lines, is sometimes used to reduce dispersion in lines.

Example 2.2.12

A transmission line has the following primary constants: R=23 Ω km^{-1}, G = 4 mS km^{-1}, L = 125 μH km^{-1} and C = 48 nF km^{-1}. Calculate the propagation constant γ of the line, and the characteristic impedance Z_0 of the line at a frequency of (a) 100 Hz, (b) 500 Hz, (c) 15 kHz, (d) 5 MHz and (e) 10 MHz.

Solution

The characteristic impedance Z_0 will not be calculated here because it has already been carried out in Example 2.2.4. However, the results will be copied to allow easy comparison with the propagation results calculated here for the discussion that follows after this answer. Equation 2.2.41 will be used to calculate the propagation constant γ, and Equation 2.2.42 will be used to derive the attenuation constant α and the phase constant β in all the calculations that follow.

(a) At 100 Hz

$$R + j\omega L = (23 + j(2\pi \times 100 \times 125\mu H)) = (23 + j0.08)\ \Omega\ km^{-1}$$

and

$$G + j\omega C = (4\ mS + j(2\pi \times 100 \times 48\ nF)) = (4 + j0.030)\ mS\ km^{-1}$$

Hence

$$\gamma = \sqrt{(23 + j0.08)(4 + j0.030) \times 10^{-3}} = 0.30\underline{/0.01} = (0.30\ nepers + 1.66 \times 10^{-3} rad)\ km^{-1}$$

and

$$Z_0 = \sqrt{\frac{23 + j0.08}{(4 + j0.030) \times 10^{-3}}} = 75.83\Omega\underline{/-2.06} \times 10^{-3}\ rad$$

(b) At 500 Hz

$$R + j\omega L = (23 + j(2\pi \times 500 \times 125\mu H)) = (23 + j0.39)\ \Omega\ km^{-1}$$

and

$$G + j\omega C = (4\ mS + j(2\pi \times 500 \times 48\ nF)) = (4 + j0.15)\ mS\ km^{-1}$$

Hence

$$\gamma = \sqrt{(23 + j0.39)(4 + j0.15) \times 10^{-3}} = 0.30\underline{/0.03}\ rad = (0.30\ nepers + 8.31 \times 10^{-3} rad)\ km^{-1}$$

and

$$Z_0 = \sqrt{\frac{23 + j0.39}{(4 + j0.15) \times 10^{-3}}} = 75.82\Omega\underline{/-10.30 \times 10^{-3}}\ rad$$

(c) At 15 kHz

$$R + j\omega L = (23 + j(2\pi \times 15 \times 10^3 \times 125\mu H)) = (23 + j11.78)\ \Omega\ km^{-1}$$

and

$$G + j\omega C = (4\ mS + j(2\pi \times 15 \times 10^3 \times 48\ nF)) = (4 + j4.52 \times 10^{-3})\ mS\ km^{-1}$$

Hence

$$\gamma = \sqrt{(23 + j11.78)(4 + j4.52) \times 10^{-3}} = 0.40\underline{/0.66}\ rad = (0.31\ nepers + 242 \times 10^{-3} rad)\ km^{-1}$$

and

$$Z_0 = \sqrt{\frac{23 + j11.78}{(4 + j4.52) \times 10^{-3}}} = 65.42\ \Omega\underline{/ -0.19}\ rad$$

(d) At 5 MHz

$$R + j\omega L = (23 + j(2\pi \times 5 \times 10^6 \times 125\mu H)) = (23 + j3926.99)\ \Omega\ km^{-1}$$

and

$$G + j\omega C = (4\ mS + j(2\pi \times 5 \times 10^6 \times 48\ nF)) = (4 + j1508)\ mS\ km^{-1}$$

Hence

$$\gamma = \sqrt{(23 + j3926.99)(4 + j1508) \times 10^{-3}} = 76.95\underline{/1.567}\ rad = (0.33\ nepers + 76.95\ rad)\ km^{-1}$$

and

$$Z_0 = \sqrt{\frac{23 + j3926.99}{(4 + j1508) \times 10^{-3}}} = 50.03\ \Omega\underline{/ -0.00}\ rad$$

(e) At 10 MHz

$$R + j\omega L = (23 + j(2\pi \times 10 \times 10^6 \times 125\mu H)) = (23 + j7853.98)\ \Omega\ km^{-1}$$

and

$$G + j\omega C = (4\ mS + j(2\pi \times 10 \times 10^6 \times 48\ nF)) = (4 + j3016)\ mS\ km^{-1}$$

Hence

$$\gamma = \sqrt{(23 + j7853.98)(4 + j3016) \times 10^{-3}} = 153.91\underline{/1.569}\ rad = (0.33\ nepers + j153.9\ rad)\ km^{-1}$$

$$Z_0 = \sqrt{\frac{23 + j7853.98}{(4 + j3016) \times 10^{-3}}} = 50.03\ \Omega\underline{/ -0.00}\ rad$$

Conclusions from Example 2.2.12. In the frequency range 100–500 Hz, the attenuation constant α tends to remain at about 0.30 nepers per km and the phase propagation β increases linearly with frequency. See cases (a) and (b). If you now compare this set of results with the same cases from Example 2.2.4, you will see that in this frequency range, Z_0 and α tend to remain constant and β tends to vary linearly with frequency.

What this means is that if you transmit a rectangular pulse or digital signals in this frequency range, you will find that it will pass through the transmission line attenuated but with its shape virtually unchanged. The reason for the waveform shape not changing is because the Fourier amplitude and phase relationships have not been changed.

In the frequency range 5–10 MHz, the attenuation constant α tends to remain at 0.33 nepers per km and the phase propagation β also increases linearly with frequency. See cases (d) and (e). If you now compare the above set of cases from Example 2.2.12 with an identical set from Example 2.2.4, you will see that within these frequency ranges, Z_0 and α tend to remain constant and β tends to vary linearly with frequency. Therefore the same argument in the foregoing paragraphs applies to this frequency range. This is also known as the 'distortionless' range of the transmission line.

In the intermediate frequency range of operation (see case (c) of Examples 2.2.4 and 2.2.12), both the propagation constant α and β, and the characteristic impedance of the line Z_0 vary. Fourier amplitude and phase relations are not maintained as waveforms are transmitted along the line and waveform distortion is the result.

2.2.12.3 Summary of propagation properties of transmission lines

There are two frequency regions where signals can be passed through transmission lines with minimum

distortion: a low frequency region and a high frequency region. The low frequency region occurs when $R \gg \omega L$ and $G \gg \omega C$. The high frequency region occurs when $\omega L \gg R$ and $\omega C \gg G$. The high frequency region is also sometimes called the 'lossless' region of transmission.

At both these high and low frequency regions of operation, the simplified expressions for Z_0 (Equations 2.2.18 and 2.2.19) and γ (Equations 2.2.43 and 2.2.44) show that there is little distortion and that the transmission line can be more easily terminated by a matched resistor.

Good cables which operate up to 50 GHz are available. They are relatively costly because of the necessary physical tolerances required in their manufacture.

2.2.13 Transmission lines as electrical components

Transmission lines can be made to behave like electrical components, such as resistors, inductors, capacitors, tuned circuits and transformers. These components are usually made by careful choice of transmission line characteristic impedance (Z_0), line length (l) and termination (Z_L). The properties of these components can be calculated by using well known expressions for calculating the input impedance of a transmission line.

2.2.13.1 Impedance relations in transmission lines

We shall now recall some transmission line properties which you learnt in Sections 2.2.11.2 and 2.2.12.2 to show you how the input impedance varies along the line and how transmission lines can be manipulated to produce capacitors, inductors, resistors and tuned circuits. These are Equations 2.2.28 and 2.2.29 which are repeated below for convenience:

$$\Gamma_v = \frac{Z_L - Z_0}{Z_L + Z_0} \tag{2.2.28}$$

and

$$\Gamma_i = -\Gamma_v \tag{2.2.29}$$

In previous derivations, voltage and current references have been taken from the input end of the line. Sometimes, it is more convenient to take voltage and current references from the terminating or load end of the line. This is shown in Figure 2.2-14.

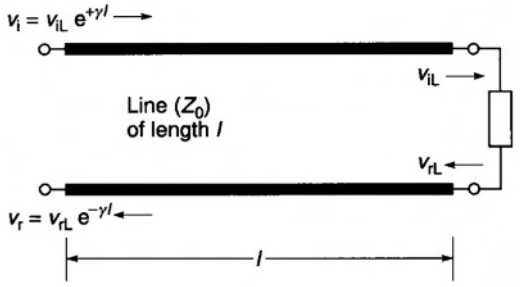


Fig. 2.2-14 Line voltages reference to the load end.

From the definition of line attenuation and for a distance l from the load, we have

$$v_i = V_{iL} e^{+\gamma l} \tag{2.2.45}$$

and

$$v_r = V_{rL} e^{-\gamma l} \tag{2.2.46}$$

and using the definition for voltage reflection coefficient Γ_v

$$\Gamma_{vl} = \frac{V_{rl}}{V_{il}} = \frac{V_{rL} e^{-\gamma l}}{V_{rL} e^{+\gamma l}} = \frac{V_{rL}}{V_{iL}} = \Gamma_L e^{-2\gamma l} \tag{2.2.47}$$

where

l = line length equal to distance l
Γ_v = voltage reflection coefficient at load
Γ_{vl} = voltage reflection coefficient at distance l from load
γ = propagation constant = $(\alpha + j\beta)$ nepers/m

At any point on a transmission line of distance l from the load

$$v_l = v_i + v_r = v_i + v_i \Gamma_v e^{-2\gamma l} \tag{2.2.48}$$

and

$$i_l = i_i + i_r = i_i + i_i \Gamma_i e^{-2\gamma l} \tag{2.2.49}$$

Dividing Equation 2.2.48 by Equation 2.2.49 and using Equations 2.2.28 and 2.2.29

$$\frac{v_l}{i_l} = \frac{v_i + v_i \Gamma_v e^{-2\gamma l}}{i_i - i_i \Gamma_v e^{-2\gamma l}} = \frac{v_i (1 + \Gamma_v e^{-2\gamma l})}{i_i (1 - \Gamma_v e^{-2\gamma l})} \tag{2.2.50}$$

Defining Z_l as impedance at point l, and Z_0 as the line characteristic impedance, Equation 2.2.50 becomes

$$Z_l = Z_0 \left[\frac{1 + \Gamma_v e^{-2\gamma l}}{1 - \Gamma_v e^{-2\gamma l}} \right] \tag{2.2.51}$$

Substituting Equation 2.2.28 in Equation 2.2.51 results in

$$Z_l = Z_0\left[\frac{1+\frac{Z_L-Z_0}{Z_L+Z_0}e^{-2\gamma l}}{1-\frac{Z_L-Z_0}{Z_L+Z_0}e^{-2\gamma l}}\right]$$

Multiplying out and simplifying

$$Z_l = Z_0\left[\frac{Z_L+Z_0+(Z_L-Z_0)e^{-2\gamma l}}{Z_L+Z_0-(Z_L-Z_0)e^{-2\gamma l}}\right]$$

Sorting out Z_0 and Z_L gives

$$Z_l = Z_0\left[\frac{Z_L(1+e^{-2\gamma l})+Z_0(1-e^{-2\gamma l})}{Z_L(1-e^{-2\gamma l})+Z_0(1+e^{-2\gamma l})}\right]$$

Multiplying all bracketed terms by $(e^{\gamma l}/2)$ results in

$$Z_l = Z_0\left[\frac{Z_0\frac{e^{\gamma l}-e^{-\gamma l}}{2}+Z_L\frac{e^{\gamma l}+e^{-\gamma l}}{2}}{Z_0\frac{e^{\gamma l}+e^{-\gamma l}}{2}+Z_L\frac{e^{\gamma l}-e^{-\gamma l}}{2}}\right]$$

Bear in mind that by definition

$$\sinh\gamma l = \frac{e^{\gamma l}-e^{-\gamma l}}{2} \quad \text{and} \quad \cosh\gamma l = \frac{e^{\gamma l}-e^{-\gamma l}}{2}$$

Substituting for $\sinh\gamma l$ and $\cosh\gamma l$ in the above equation results in

$$Z_l = Z_0\left[\frac{Z_0\sinh\gamma l+Z_L\cosh\gamma l}{Z_0\cosh\gamma l+Z_L\sinh\gamma l}\right] \tag{2.2.52}$$

Equation 2.2.52 is a very important equation because it enables us to investigate the properties of a transmission line. If the total length of the line is l then Z_1 becomes the input impedance of the line. Hence, Equation 2.2.52 becomes

$$Z_{in} = Z_0\left[\frac{Z_0\sinh\gamma l+Z_L\cosh\gamma l}{Z_0\cosh\gamma l+Z_L\sinh\gamma l}\right] \tag{2.2.53}$$

2.2.13.2 Input impedance of low loss transmission lines

From Equation 2.2.42, we know that $\gamma = \alpha + j\beta$. When $\alpha \ll \beta$, $\gamma = j\beta$.

From Equation 2.2.22, we know that $\beta = 2\pi/\lambda$. From mathematical tables[17] we know that $\sin(j\beta) = j\sin\beta$ and $\cos(j\beta) = \cos\beta$. If we now substitute the above facts into Equation 2.2.53, we will get Equation 2.2.54. The input impedance of a low loss transmission line is given by the expression

$$Z_{in} = Z_0\left[\frac{jZ_0\sin\frac{2\pi l}{\lambda}+Z_L\cos\frac{2\pi l}{\lambda}}{jZ_L\sin\frac{2\pi l}{\lambda}+Z_0\cos\frac{2\pi l}{\lambda}}\right] \tag{2.2.54}$$

where

Z_{in} = input impedance (ohms)
Z_0 = characteristic impedance of line (ohms)
Z_L = termination load on line (ohms)
λ = electrical wavelength at the operating frequency
l = transmission line length

2.2.13.3 Reactances using transmission lines

A transmission line can be made to behave like a reactance by making the terminating load a short circuit ($Z_L = 0$). In this case, Equation 2.2.54 becomes

$$Z_{in} = Z_0\left[\frac{jZ_0\sin\frac{2\pi l}{\lambda}+0}{0+Z_0\cos\frac{2\pi l}{\lambda}}\right] = Z_0\left[\frac{j\sin\frac{2\pi l}{\lambda}}{\cos\frac{2\pi l}{\lambda}}\right]$$
$$= jZ_0\tan\frac{2\pi l}{\lambda} \tag{2.2.55}$$

When $l < \lambda/4$

$$Z_{in} = jZ_0\tan\frac{2\pi l}{\lambda} \tag{2.2.55a}$$

and is inductive.

When $\lambda/4 < l > \lambda/2$

$$Z_{in} = -jZ_0\tan\frac{2\pi l}{\lambda} \tag{2.2.55b}$$

and is capacitive.

Equation 2.2.55 follows a tangent curve and like any tangent curve it will yield positive and negative values. Therefore Equation 2.2.55 can be used to calculate inductive and capacitive reactances. Adjustment of Z_0 and line length, l, will no doubt set the required values.

Example 2.2.13

A 377 Ω transmission line is terminated by a short circuit at one end. Its electrical length is $\lambda/7$. Calculate its input impedance at the other end.

[17] You can also check this for yourself if you take the series for sin x and cos x and substitute $j\beta$ in place of x.

Solution

Using Equation 2.2.55a

$$Z_{in} = jZ_0 \tan\frac{2\pi l}{\lambda} = j377 \tan\left[\frac{2\pi}{\lambda}\frac{\lambda}{7}\right]$$
$$= j377 \times 1.254 = j472.8\ \Omega$$

Similar reactive effects can also be produced by using an open-circuited load[18] and applying it to Equation 2.2.54 to produce inductive and capacitive reactances:

$$Z_{in} = -jZ_0 \cot\frac{2\pi l}{\lambda} \tag{2.2.56}$$

Equation 2.2.56 follows a cotangent curve and will therefore also produce positive and negative impedances. Adjustment of Z_0 and line length will set the required reactance.

Example 2.2.14

A 75 Ω line is left unterminated with an open circuit at one end. Its electrical length is $\lambda/5$. Calculate its input impedance at the other end.

Solution

Using Equation 2.2.56

$$Z_{in} = -jZ_0 \cot\frac{2\pi l}{\lambda} = -j75 \cot\left[\frac{2\pi}{\lambda}\frac{\lambda}{5}\right]$$
$$= -j75 \times 0.325 = -j24.4\ \Omega$$

2.2.13.4 Transmission lines as transformers

An interesting case arises when $l = \lambda/2$. In this case Equation 2.2.54 becomes

$$Z_{in} = Z_0\left[\frac{jZ_0 \sin\pi + Z_L \cos\pi}{jZ_L \sin\pi + Z_0 \cos\pi}\right] = Z_L$$

What this means is that the transmission line acts as a 1:1 transformer which is very useful for transferring the electrical loading effect of a termination which cannot be placed in a particular physical position. For example, a resistor dissipating a lot of heat adjacent to a transistor can cause the latter to malfunction. With a 1:1 transformer, the resistor can be physically moved away from the transistor without upsetting electrical operating conditions.

Another interesting case arises when $l = \lambda/4$. In this case, Equation 2.2.54 becomes

$$Z_{in} = Z_0\left[\frac{jZ_0 \sin\frac{\pi}{2} + Z_L \cos\frac{\pi}{2}}{jZ_L \sin\frac{\pi}{2} + Z_0 \cos\frac{\pi}{2}}\right] = \frac{Z_0^2}{Z_L}$$

Therefore the transmission line behaves like a transformer where

$$Z_{in} = \frac{Z_0^2}{Z_L} \tag{2.2.57}$$

At first glance Equation 2.2.57 may not seem to be very useful but if you refer back to Figure 2.2-7, you will see that the characteristic impedance (Z_0) of microstrip transmission lines can be easily changed by changing its width (w); therefore, impedance matching is a very practical proposition.

Example 2.2.15

A transmission line has a characteristic impedance (Z_0) of 90 Ω. Its electrical length is $\lambda/4$ and it is terminated by a load impedance (Z_L) of 20 Ω. Calculate the input impedance (Z_{in}) presented by the line.

Given: $Z_0 = 90\ \Omega$, $Z_L = 20\ \Omega$, $l = \lambda/4$
Required: Z_{in}

Solution

Using Equation 2.2.57

$$Z_{in} = (90)^2/20 = 405\ \Omega$$

2.2.14 Transmission line couplers

Transmission lines can be arranged in special configurations to divide an input signal at the input port into two separate signals at the output ports. Such an arrangement is often called a signal splitter. Since the splitter is bidirectional, the same arrangement can also be used to combine two separate signals into one. These splitter/combiners are often called couplers.

The advantages of couplers are that they are very efficient (low loss), provide good matching on all ports, and offer reasonable isolation between the output ports so that one port does not interfere with the other. The greatest disadvantage of these couplers is their large physical size when used in their distributed forms.

[18] Purists might argue that an open circuit does not exist at radio frequencies because any unterminated TX line has stray capacitance associated with an open circuit. We will ignore this stray capacitance temporarily because, for our frequencies of operation, its reactance is extremely high.

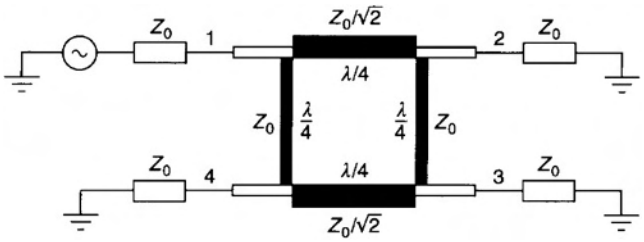


Fig. 2.2-15 A branch-line coupler.

2.2.14.1 The branch-line coupler

The basic configuration of the branch-line coupler is shown in Figure 2.2-15. It consists of transmission lines, each having a length of $\lambda/4$. Two opposite facing-lines have equal impedances, Z_0, and the remaining opposite facing-lines have an impedance of $Z_0/\sqrt{2}$. In the case of a 50 Ω coupler, $Z_0 = 50\ \Omega$ and $Z_0/\sqrt{2} = 35.355\ \Omega$.

Principle of operation

All ports are terminated in Z_0. Signal is applied to port 1 and it is divided equally between ports 2 and 3. There is no output signal from port 4 because the path from port 1 to 4 is $\lambda/4$ while the path from port 1 to 4 via ports 2 and 3 is $3\lambda/4$. The path difference is $\lambda/2$; hence, the signals cancel each other at port 4. The net result is that the signal at port 4 is zero and it can be considered as a virtual earth point. With this virtual earth point Figure 2.2-15 becomes Figure 2.2-16(a). We know from transmission line theory (Equation 2.2.57) that for a $\lambda/4$ length, $Z_{in} = Z_0^2/Z_l$. In other words, a short circuit at port 4 appears as open circuits at port 1 and 3. This result is shown in Figure 2.2-16(b). If we now transform the impedance at port 3 to port 2, we get

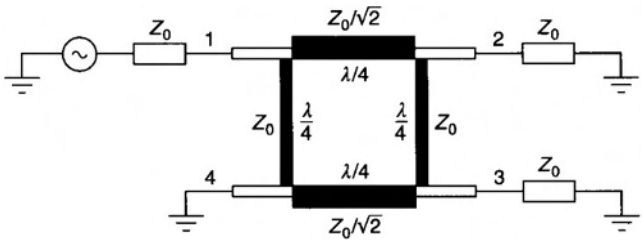


Fig. 2.2-16(a) Virtual short circuit at port 4.

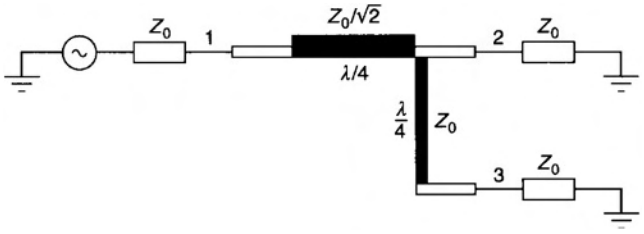


Fig. 2.2-16(b) Effect of virtual short circuit at port 4.

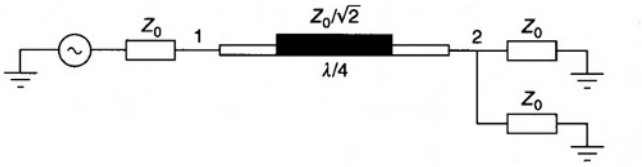


Fig. 2.2-16(c) Effect of port 3 transferred to port 2.

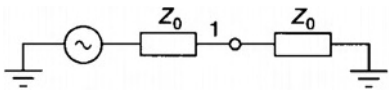


Fig. 2.2-16(d) Effect of port 2 transferred to port 1.

$$Z_{(\text{transformed})} = Z_0^2/Z_0 = Z_0$$

Hence we obtained the transformed Z_0 in parallel with the Z_0 termination of port 2. This situation is shown in Figure 2.2-16(c). We now need to transform the $Z_0/2$ termination at port 2 to port 1. At port 1

$$Z_l = \frac{(Z_0/\sqrt{2})^2}{(Z_0/2)} = \frac{Z_0^2}{2} \times \frac{2}{Z_0} = Z_0$$

This condition is shown in Figure 2.2-16(d).

We conclude that the hybrid 3 dB coupler provides a good match to its source impedance Z_0. Assuming lossless lines, it divides its signal equally at ports 2 and 3. Since signal travels over $\lambda/4$ to port 2, there is a phase delay of 90° at port 2 and another 90° phase delay from port 2 to port 3. Thus the signal arriving at port 3 suffers a delay of 180° from the signal at port 1.

The response of such a coupler designed for 5 GHz is shown in Figure 2.2-17. As you can see for yourself, the signal path from port 1 to port 2 (S_{21}) is 3 dB down at 5 GHz. This is also true of the signal response to port 3 (S_{31}). The signal attenuation to port 4 is theoretically infinite but this value is outside the range of the graph.

A similar analysis will show that power entering at port 2 will be distributed between ports 1 and 4 and not at port 3. A similar analysis for port 3 will show power dividing between ports 1 and 4 but not at port 2. The net result of this analysis shows that signals into port 2 and 3 are isolated from each other. This is a very useful feature in mixer and amplifier designs.

The advantage of the quadrature coupler is easy construction but the disadvantage of this coupler is its narrow operational bandwidth because perfect match is only obtained at the design frequency where each line is exactly λ/4 long. At other frequencies, each line length is no longer $\lambda/4$ and signal attenuation increases while signal isolation decreases between the relevant ports.

Finally, you may well ask 'if port 4 is at virtual earth, why is it necessary to have a matched resistor at port 4?'

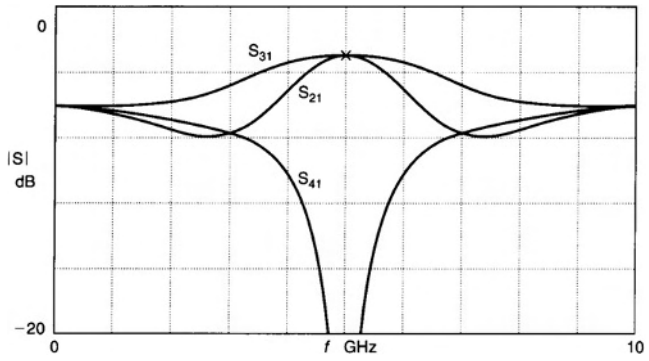


Fig. 2.2-17 Unadjusted response of the quadrate 3 dB coupler: S_{21} = signal attenuation path from port 1 to port 2; S_{31} = signal attenuation path from port 1 to port 3; S_{41} = signal attenuation path from port 1 to port 4.

The reason is that signal balance is not perfect and the resistor helps to absorb unbalanced signals and minimises reflections.

2.2.14.2 The ring coupler

Ring forms of couplers have been known for many years in waveguide, coaxial and stripline configurations. The basic design requirements are similar to that of the quadrature coupler except that curved lines are used instead of straight lines. One such coupler is shown in Figure 2.2-18. The principle of operation of ring couplers is similar to that of branch-line or quadrature couplers.

2.2.14.3 The 'rat-race' coupler

A sketch of a 'rat-race' coupler is shown in Figure 2.2-19. The mean circumference of the ring is 1.5λ. This coupler is easy to construct and provides good performance for narrow band frequencies. The characteristic impedance of the coupler ring is $Z_0\sqrt{2}$ Ω which in the case of $Z_0 = 50$ Ω is a circular 70.7 Ω transmission line which is 1.5λ, in circumference. The four Z_0 ports are connected to the ring in such a manner that ports 2 to 3, ports 3 to 1 and ports 1 to 4 are each separated by λ/4. Port 4 to port 2 is separated by 0.75λ. The operation of this device is illustrated in Figure 2.2-20.

If a signal is injected at port 1, the voltage appearing at port 2 is zero, since the path lengths differ by 0.5λ; thus, port 2 can be treated as a virtual ground. Hence, the transmission-line portions of the ring between ports 2

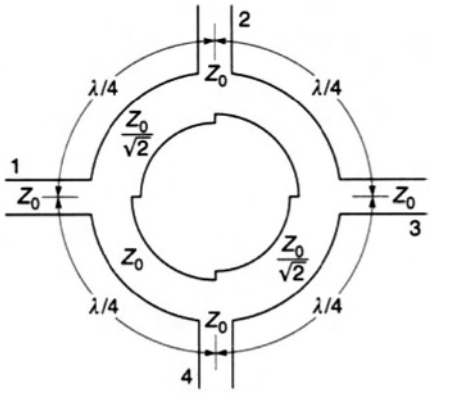


Fig. 2.2-18 The 3 dB ring-form branch-line directional coupler.

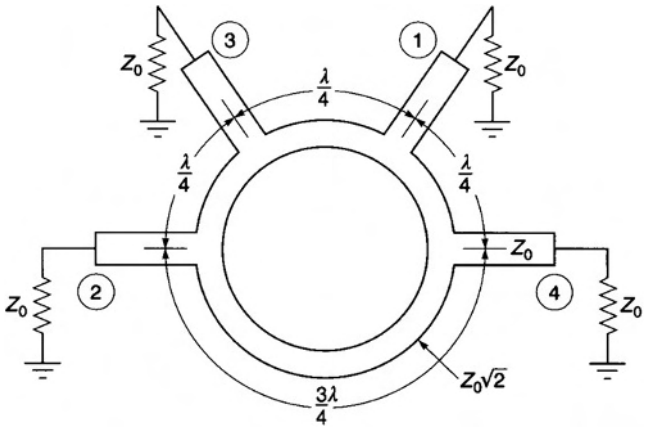


Fig. 2.2-19 The 'rat-race' or 'hybrid ring' coupler.

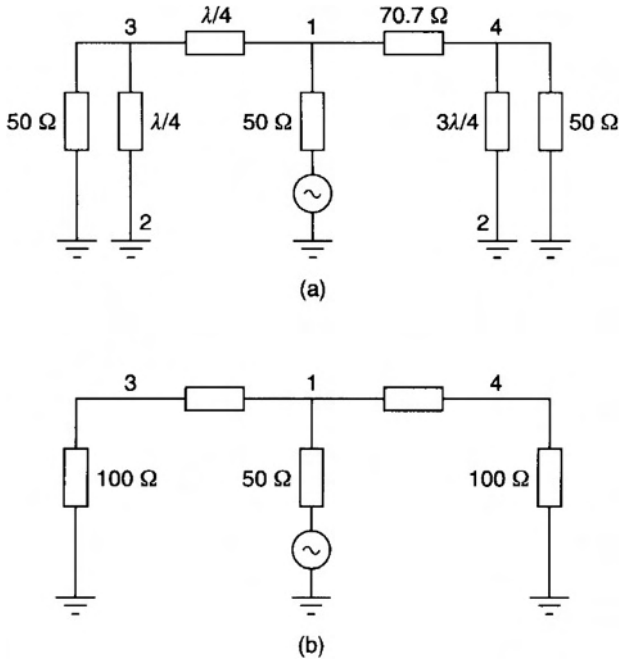


Fig. 2.2-20 (a) Equivalent circuit of ring hybrid with port 1 as input and ports 2 and 4 as outputs (transmission-line model with port 3 as virtual ground); (b) equivalent circuit at centre frequency.

and 3, and ports 2 and 4, act as short-circuited stubs connected across the loads presented at ports 3 and 4. For centre frequency operation, these stubs appear as open circuits. Similarly, the transmission line lengths between ports 3 and 1, and ports 4 and 1, transform the 50 Ω load impedances at ports 3 and 4 to 100 Ω ($2Z_0$) at port 1. When combined at port 1, these transformed impedances produce the 50 Ω impedance at port 1 (see Figures 2.2-20(a) and (b)).

A similar analysis can be applied at each port, showing that the hybrid exhibits a matched impedance of 50 Ω or Z_0 at all nodes. It should be noted that when port 1 is driven, the outputs at ports 3 and 4 are equal and in phase, while ideally there is no signal at port 2. However, when port 2 is driven, the output signals appearing at ports 3 and 4 are equal but exactly out of phase. Also there is no signal at port 1. Hence ports 1 and 2 are isolated. This is very useful especially in signal mixing circuits because it enables two slightly different frequencies, for example f_1 at port 1 and f_2 at port 2, to be applied to a balanced mixer whose diodes may be connected to ports 3 and 4 without coupling between the sources at port 1 and port 3. It also helps to combine the inputs or outputs of two amplifiers without mutual interference. The unfortunate thing about the ring is that it is a relatively narrow-band device.

2.2.15 Summary

This chapter has provided you with a thorough basic knowledge of transmission lines and their properties which you will find very useful in circuit and system design of radio and microwave systems. You have been introduced to many properties of transmission lines in this chapter.

In Section 2.2.3, you were introduced to some of the more frequently used types of transmission lines. These included waveguides, coplanar waveguides, coaxial lines, microstrip and strip lines, slot lines, twin lines and finally coupled microstrip lines.

In Section 2.2.4, you were shown how the characteristic impedance of the coaxial, twin line and microstrip line can be calculated from its physical parameters. The information demonstrated what properties you should look for if the characteristic impedance of a line does not behave as expected.

Sections 2.2.5 and 2.2.6 demonstrated how the characteristic impedance can be calculated and measured from primary constants of the line. Section 2.2.7 mentions some of the more common impedances associated with commercial transmission lines but it also brought to your attention that there are many types of lines with the same impedance.

Section 2.2.8 explained how propagation delay, attenuation and frequency dispersion affect waveforms as they travel along transmission lines. This was followed by more discussions on the effects of these properties in Section 2.2.9.

In Section 2.2.10, we introduced the concepts of matched and unmatched lines. This was followed by a thorough discussion of reflection coefficients and VSWRs in Section 2.2.11.

Section 2.2.12 dealt with the propagation properties of lines and showed how these can be derived from the primary constants of the line. The section also showed how optimum transmission can be achieved.

Section 2.2.13 showed how transmission lines can be used as transformers, impedance matching devices, inductive and capacitive reactances which in turn can be used to produce filters. Microstrip lines are particularly useful for making filters at the higher frequencies because of their versatility in allowing characteristic impedance changes to be made easily. There is also a greater tendency to use transmission lines as the tuning elements.

Section 2.2.14 showed you how transmission lines can be connected to act as signal couplers.

This page intentionally left blank

Chapter 2.3

2.3

Software defined radio

Brad Bannon

What Is software-defined radio?

Over the last decade as semiconductor technology has improved both in terms of performance capability and cost, new radio technologies have emerged from military and R&D labs and become mainstream technologies. One of these technologies is software-defined radio (SDR). Although much has been discussed in recent years, a good definition of software radio is difficult to generate. This is largely due to the flexibility that SDRs offer, allowing them to take on many different forms that can be changed to suit the need at hand.

However, SDRs do have characteristics that make them unique in comparison to other types of radios. As the name implies, an SDR is a radio that has the ability to be transformed through the use of software or redefinable logic. Quite often this is done with general purpose digital signal processors (DSPs) or field programmable gate arrays (FPGAs), as discussed later in the chapter. In order to take advantage of such digital processing, traditional analog signals must be converted to and from the digital domain. This is accomplished using analog-to-digital (ADC) and digital-to-analog converters (DAC). To take full advantage of digital processing, SDRs keep the signal in the digital domain for as much of the signal chain as possible, digitizing and reconstructing as close to the antenna as possible, which allows digital techniques to perform functions traditionally done by analog components as well as others not possible in the analog domain. There are limits to this, however. Despite the fact that an ADC or DAC connected directly to an antenna is a desirable end goal, there are issues with selectivity and sensitivity that an analog front end (AFE) can remedy. The alternative to digitizing at the antenna is the use of a completely flexible AFE capable of translating a wide range of frequencies and bands to that which the data converters themselves can adequately process [1].

SDRs are ideal candidates to be used for multi-carrier, single-carrier, single-band, multiband, and multimode transceivers. Some of these issues will be covered later. The key point is that SDRs have the ability to go beyond simple single-channel, single-mode transceiver technology with the ability to change modes arbitrarily because the channel bandwidth, rate, and modulation are all flexibly determined through software. These characteristics may be changed by direct input, floppy disk, over-the-air download, or through the use of careful signal analysis to determine analytically how the information is coded through a process termed *cognitive radio* [2]. Regardless of the means by which the radio is reconfigured, a fully implemented SDR will have the ability to navigate a wide range of frequencies with programmable channel bandwidth and modulation characteristics. The following list outlines some of the possible dynamic characteristics of an SDR:

- Channel bandwidth
- Data rate
- Modulation type
- Conversion gain

In addition to RF tuning, a transceiver must include the ability to take advantage of one or more of these characteristics to be considered an SDR.

Aspects of SDR

As the preceding list indicates, there are a number of characteristics that an SDR possesses. Although it is not

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2004 Elsevier Inc. All rights reserved

required that an SDR have all of these characteristics, having one or more of them is. Additionally, the categories above can be further broken down as detailed in the following sections. It should be kept in mind that since software-defined implies a high degree of flexibility and variability, this discussion is not all encompassing and is subject to change over time. Nevertheless, it will serve as a starting point for understanding the different facets of what SDR can be.

Multiband

Most traditional radio architectures operate on a single band or range of frequencies. There are many applications where multiple frequencies of operations are desired. These include cellular communications, government and nongovernment agencies, and intelligence collection, to list a few. Where these situations exist, the norm is to utilize multiple radios, each designed to operate in one specified band. A multiband radio has the ability to operate on two or more bands either sequentially or simultaneously, as in the case of a base station that may be linking handsets from different bands.

Multi-carrier

A multi-carrier or multi-channel radio has the ability to simultaneously operate on more than one frequency at a time. This may be within the same band or, in the case of a multiband radio, in two different bands at the same time. Quite often, multi-carrier applies to a base station that may be servicing many users at once, but it can also apply to a user terminal that may be processing both voice and data on different RF carriers.

Multimode

Multimode implies the ability to process several different kinds of standards. Examples of standards include AM, FM, GMSK, and CDMA but are limited to none of these. An SDR has the ability to work with many different standards and be continuously reprogrammed. Therefore, a better term than *multimode*, which implies a discrete number of modes, may be *variable mode*, which implies a continuously changeable mode of operation. As with other characteristics, these modes may be sequential or simultaneous, in the case of a multi-carrier radio.

Multi-rate

Multi-rate is closely related to multimode. A multi-rate radio is one that either processes different parts of the signal chain at different samples rates, as in a multi-rate filter, or one where the radio has the ability to process different modes that require different data rates. An example of a multi-rate radio would be one that can process GSM at 270.833 kSPS or CDMA at 1.2288 MCPS. As with other characteristics, this can be done sequentially or at the same time on different carriers.

Variable bandwidth

Variable bandwidth is another aspect of multimode. A traditional radio determines the channel bandwidth with a fixed analog filter such as an SAW or ceramic filter. A SDR, however, determines the channel bandwidth using digital filters that can be altered. While a series of switched analog filters could be used to change the channel bandwidth in a traditional receiver, only a small number would be practical. Additionally, digital filters have the potential to implement filters not possible in the analog domain. Lastly, digital filters can be tailored to both adapt around interferers and compensate for transmission path distortion, both features that analog filters are hard pressed to accomplish.

History and evolution of software-defined radio

The history of SDR began in the mid-1980s. One of the first major developments for SDR was the SpeakEasy, a transceiver platform designed by Hazeltine and Motorola that was based on SDR technology for Rome Air Force Base, New York. The SpeakEasy was designed to provide tactical military communications from 2 MHz to 2 GHz and to provide interoperability between the different air interface standards of the different branches of the armed forces. To achieve this goal, the SpeakEasy utilized many of the techniques discussed in this chapter to provide multiband, multi-modes of operations. Although many people contributed to the concept and development of SDR, Joe Mitola of Mitre is generally credited with being the "father of software-defined radio" [2].

Although the SpeakEasy was a fully developed SDR, it is fair to say that simpler and more rudimentary forms of SDR existed before this program. By taking a look at how systems are being developed in the commercial realm, it is easy to see how they also may have evolved in military and nonmilitary programs.

Although there are many enabling technologies that have come online in the last decade, one of the key technical driving forces was the development of low-cost DSPs. From a market point of view, the rapid growth of the telecommunications industry, particularly cellular communications, provided a demand for low-cost

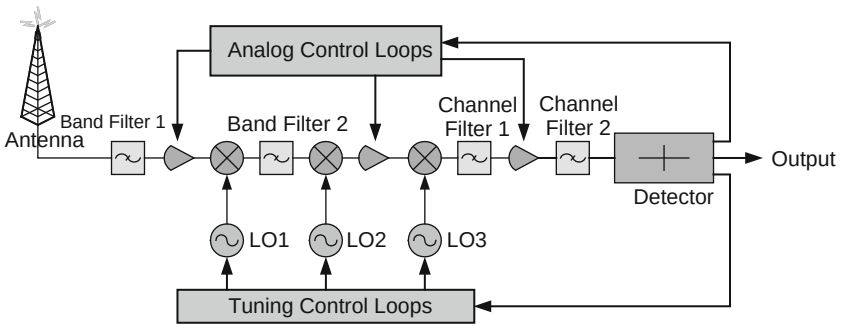


Figure 2.3-1 Traditional super-heterodyne with analog detection.

equipment both from a user and an infrastructure point of view. Although first-generation (1G) cellular was based on analog modulation schemes (which did not require significant digital processing), it became clear that due to the limited amount of spectrum and the relative inefficiency of those standards, more efficient means of spectral usage were required. Therefore, second generation (2G) cellular systems such as GSM and IS-95 were developed that took advantage of the emerging DSP technologies. In these early systems, the DSP became the *modem* function and was responsible for taking the complex baseband data (I and Q) and determining what bitstream was being sent and correcting for errors introduced due to noise, interference, and fading.

Conceptually, these modem functions were based on programs running on a DSP and, therefore, could be changed simply by changing the program. In fact, over time these standards evolved and variations of the standards were introduced that allowed better efficiency and higher data-transmission rates. Many of these improvements were offered simply by updating the modem software. While consumers seldom experienced these benefits for a number of economic reasons, the infrastructure side did benefit from these upgrades and is benefiting from many of these software updates in the migration from 2G to 2.5G and ultimately to 3G, most notably in the evolution of the CDMA2000 and UMTS standards [15, 16].

While the evolution of the modems used for GSM and CDMA is an aspect of SDR, other factors such as incompatibility of these two standards drives the second aspect of SDR. While CDMA is primarily a North American and Asian standard and GSM is a standard in Europe and the rest of world, in reality both of these standards are overlayed in many countries. Ideally, service providers would like to purchase one piece of equipment that would work with both standards. Unfortunately, these and other standards are incompatible in terms of bandwidth, modulation format, and data rate. Traditional radios, even those with DSP modems, operate with fixed bandwidths and therefore prevent crossfunctionality. A typical GSM system works with a 200 kHz bandwidth, while an IS-95 system operates on 1.25 MHz bandwidth. Both systems typically utilize surface acoustic wave (SAW) filter technology to set the bandwidth. Since these devices are fixed, it is not possible (aside from electronically switching filters) to change the channel bandwidth characteristics. Therefore, aside from the modem function, an SDR needs additional circuitry that allows other properties of the air interface to be adapted. In practice, the optimal way to do this is to digitize the signal and use digital techniques for manipulating the channel of interest. This manipulation often occurs in a general purpose DSP or more frequently in a digital ASIC with the ability to accommodate a near continuous range of channel bandwidth, data rates, and other physical characteristics in a fully digital manner.

The following figures show the practical evolution path of an Rx SDR architecture. In Figure 2.3-1, a traditional super-heterodyne receiver is shown with analog detection. Figure 2.3-2 adds the DSP function that operates in place of the analog detector and can also function as the baseband modem. This allows the exact demodulation functions to change as needed while the other channel characteristics are fixed. Figure 2.3-3 includes a wideband ADC and digital preprocessing before the modem that allows the physical channel characteristics to be adapted as necessary. Figure 2.3-3 also shows what a full SDR might look like. An architecture such as this may find use in diverse areas such as multi-mode systems capable of simultaneously processing several standards at once or simply as a manufacturing option to simplify production and inventory issues.

Applications and need for SDR

Interoperability

The military is not the only agency in need of interoperability. As numerous agencies, both domestic and international, have responded to various natural and manmade disasters around the world, communications

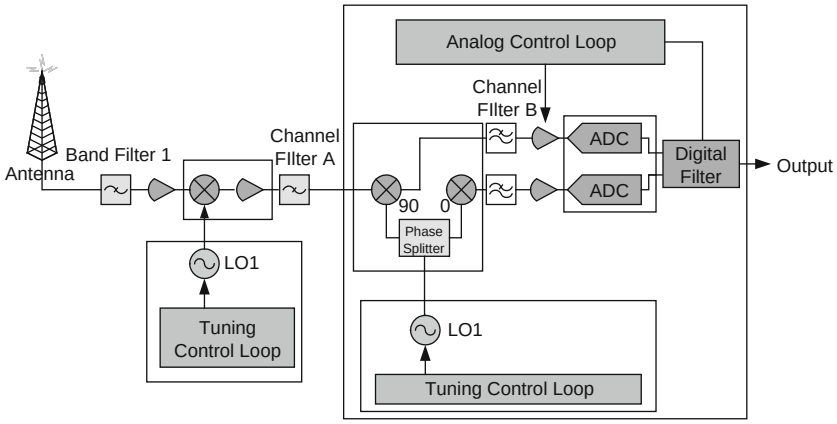


Figure 2.3-2 Super-heterodyne with baseband IQ sampling.

between the different responding groups has often been hindered by the fact that different communications systems rarely work with one another because the frequency and air interfaces are different. As shown in Figure 2.3-4, SDR provides an ideal solution to these dilemmas. A centrally deployed base station could be used to receive the transmissions of one agency and reformat and rebroadcast them on the frequencies of the other responding agencies. Since the system would be reconfigurable as new agencies arrive or depart, the SDR could be rapidly changed to support the required services. When the disaster is over, the system can easily be stowed and redeployed at a later time as required.

Interception

As already outlined, early applications of SDR were for military interoperability. Another military application for SDR is the interception of communications. Since the frequency and modulation format of these transmissions are often unknown, a flexible receiver platform capable of rapid self-adjustment is a benefit. Since SDRs can rapidly be reconfigured, they are ideal for the interception of wireless communications [3]. Additionally, since they already employ high-speed DSP, the DSP can also be utilized for advanced interception functions such as voice recognition and code decryption. Additionally, if a network of SDRs is used, then triangulation can be used to aid in the location of the rogue transmission.

Manufacturing platform

Although there are many applications where the dynamic configuration of an SDR is required, perhaps one of the most practical applications is that of a standardized communications platform. For example, most manufacturers of cellular infrastructure equipment sell platforms for a variety of air standards such as GSM, CMDA, IS-136, and AMPS, to name but a few. Typically, each of these is a different piece of hardware that must be manufactured and inventoried. If a single design could be fabricated that could have identical hardware, the cost of manufacturing would be significantly reduced because only one system would need to be inventoried. The hardware could then be configured prior to shipment or in the field for the air interface required. Although it may not be practical to look to standards of the past in and of themselves, systems of the future are prime candidates. With the competing 3G standards a manufacturer with limited resources could build a single system capable of

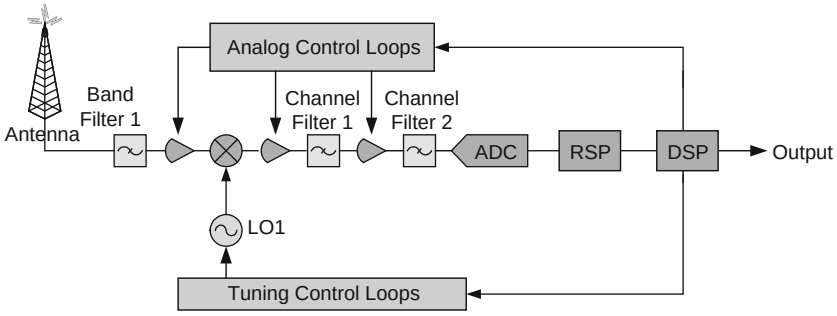


Figure 2.3-3 Super-heterodyne SDR.

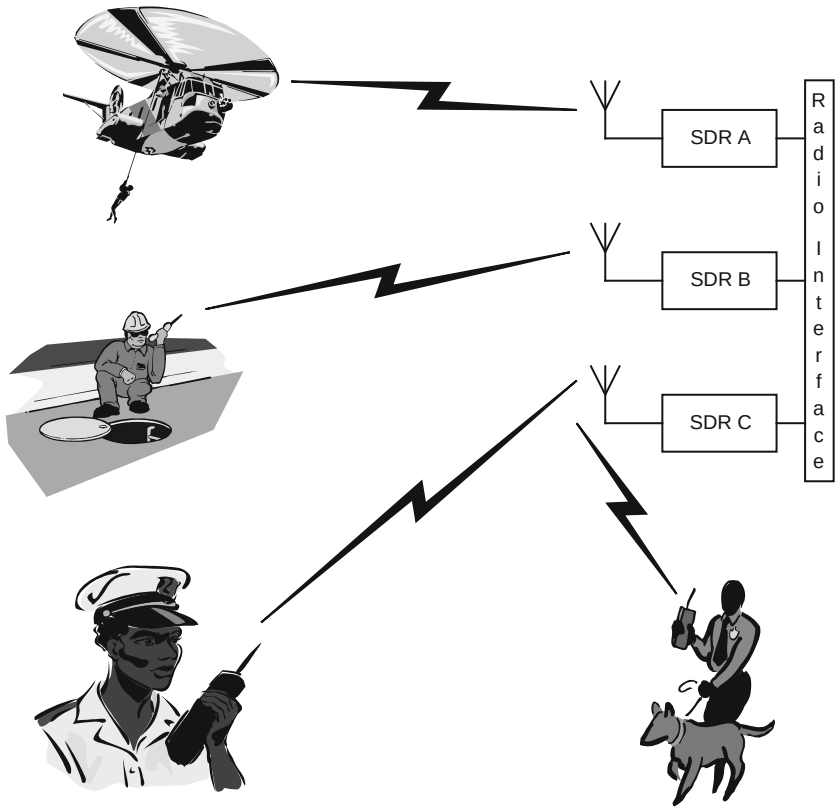


Figure 2.3-4 Interoperability.

supporting either CDMA2000 or UMTS while continuing to support the legacy standards from which these have evolved. From a user point of view, such a system is also valuable because if the user should want to change standards, all that is required is that the system be reprogrammed for the new standard, preserving all of the investment made in the original equipment.

Of course, other areas could also benefit from the economies of scale offered here. Other such areas include devices for the reception of competing satellite broadcast of audio and video content and two-way communications systems, to name a few [4].

Architectures

Ideally the designer of an SDR would like to put the data converters directly on the antenna. However, as stated previously, this is not a practical solution. In reality, some AFE must be used—before the ADC in the receive path and after the DAC in the transmit path—that does the appropriate frequency translation. The most common of these architectures is the super-heterodyne architecture. Although many decades old, new semiconductor technology and high levels of integration have kept this architecture vitalized and in popular use both in the transmit and receive signal paths [5,6]. Other architectures such as direct conversion both for transmit and receive are seeing some popularity in applications that are not as demanding. Currently direct conversion (Tx and Rx) is found in user terminals for cellular communications as well as for Tx on the base station side. It is possible that future developments will enable direct conversion on the receive side as well. Until then, the super-heterodyne architecture will continue to be used in one form or another.

Receiver

High-performance SDR receivers are typically constructed from some variant of the super-heterodyne architecture. A super-heterodyne receiver offers consistent performance across a large range of frequencies while maintaining good sensitivity and selectivity [7, 8]. Although not trivial to design, the possibility of combining wideband analog techniques and multiple front ends would allow operation across different RF bands. In the case of multi-carrier applications, this could be done simultaneously if necessary.

Multi-carrier

Depending on the applications, one or more receive channels may be desired. Traditional applications may

only require a single RF channel. However, applications that require high capacity or interoperability may require a multi-carrier design. SDRs are well suited for multi-carrier applications since they employ a highly oversampled ADC with ample available bandwidth. An oversampled ADC is one in which the sample rate is operating beyond that which is required to meet the Nyquist criterion [18], which states that the converter sample rate must be twice that of the information bandwidth. Since an SDR may not have advance knowledge of the bandwidth of the signal it will be used to receive, the sample rate must be appropriately high enough to sample all anticipated bandwidths.

Current ADC technology allows high dynamic range bandwidths of up to 100 MHz to be digitized. With this much bandwidth, it is also possible to process multiple channels. Figure 2.3-5 shows a typical multi-carrier receiver example and Figure 2.3-6 shows a spectral display. In this example, the sample rate of the ADC is set to 61.44 mega-samples-per-second (MSPS), which gives a Nyquist bandwidth of 30.72 MHz. If each RF channel is 1.25 MHz wide, then Nyquist indicates that the number of potential channel is about 24.5. In practice, by allowing for reasonable transition bands on the anti-aliasing filters, the typical available bandwidth is one-third the sample rate instead of the Nyquist one-half. Thus the available bandwidth for our example is 20.48 MHz, which is just over 16 channels at 1.25 MHz.

Since the channel characteristics can be changed, it is easy enough to change the CDMA example to a GSM example.

In this case, both the digital preprocessing and the general purpose DSP are reconfigured, respectively, by changing the digital channel filter from GSM to CDMA and by loading the new processing code into the DSP. Since GSM channels are 200 kHz wide, this example could easily be reconfigured as a 102-channel GSM receiver.

While both such examples would provide a lot of utility, perhaps a more interesting example would be to configure the receiver such that part of the channels could be CDMA while the other would be configured as GSM! Furthermore, if one of the configurations is at capacity and the other is underutilized, CDMA channels could be converted into several GSM channels or vice versa, providing the flexibility to dynamically reallocate system resources on an as-needed basis—a key goal of SDR.

Single carrier

Not all SDR applications require more than one channel. Low-capacity systems may require only one carrier. In these applications, a high over-sampling is still desired. If the channel is reprogrammable, it is possible that it may be as narrow as a few kHz or as wide as 5 to 10 MHz. In order to accommodate this range of bandwidths, the sample rate should be suitable for the highest potential bandwidth, in this case 10 MHz. From the multi-carrier example, we would typically sample at least three times the bandwidth. In this example, a sample rate of 30.72 MSPS or higher would allow signal bandwidths from a few kHz up to 10 MHz to be processed. Aside from the fact that only one channel is processed, the single-carrier receiver has all of the capacities of that of a multi-carrier receiver; it can be reconfigured as necessary.

SDR receiver elements

Referring to the single-carrier block diagram in Figure 2.3-7, while keeping in mind that this applies to the multi-carrier example as well, a fully developed SDR will have all signal elements that are programmable.

The antenna is no exception, and unfortunately it is one of the weakest elements in an SDR [1]. Since most antenna structures have a bandwidth that is a small percentage of its center frequency, multiband operation

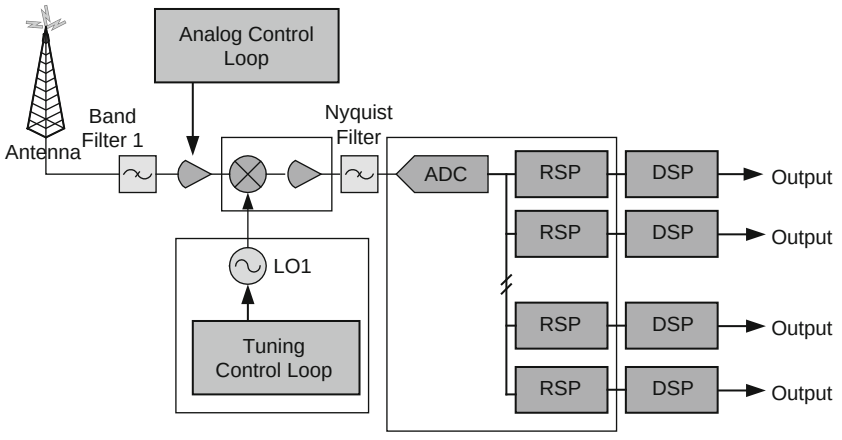


Figure 2.3-5 Multi-carrier CDMA example.

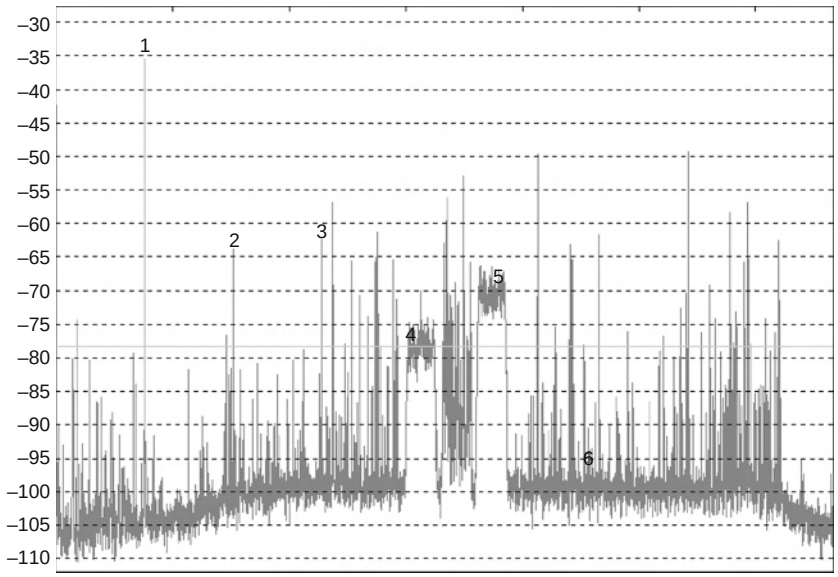


Figure 2.3-6 Multi-mode spectrum with IS-95 and narrowband carriers.

can become difficult. In the many applications where single bands of operation are used, this is not a problem. However, for systems that must operate across several orders of frequencies such as the SpeakEasy discussed earlier, the antenna must be tuned by some means to track the operating frequency to maintain operating efficiency. While it is true that just about any antenna can be impedance matched to the active electronics, there is usually a sacrifice in the link gain that potentially results in an antenna loss, whereas most antenna designs should actually provide a modest signal gain. Therefore, tuning the electrical length of the antenna is desired over simply changing the matching of the antenna.

Next in the signal chain is the band-select filter electronics. This element is provided to limit the range of input frequencies presented to the high-gain stage to minimize the effects of intermodulation distortion. Even in the case where intermodulation is not a problem, it is possible that strong out-of-band signals could limit the amount of potential gain in the following stages, resulting in limited sensitivity. This is especially true for receivers tuned near television and audio broadcast services where transmit power levels can exceed 100 kW. This can be especially problematic for multi-carrier receivers where many orders of signal magnitude must be dealt with. If all of the signals are of interest, then it will not be possible to filter the stronger signals, and the resulting receiver must have a relatively large signal dynamic range [8].

Most receivers require a low-noise amplifier or LNA. An SDR should ideally incorporate an LNA that is capable of operating over the desired range of frequencies. In addition to the typical low NF and high IP3, it may be desirable to have the ability to adjust the gain and potentially scale the power down (often NF and IP3 track bias current) when possible, which will allow for a variety of signal conditions that exist across the bands of operation.

Mixers are used to translate the RF spectrum to a suitable IF frequency. Although only one mixer is shown in Figure 2.3-7, many receivers may use two or three mixer stages, each successively generating a lower frequency. (Note: Receiver IFs are not always lower than

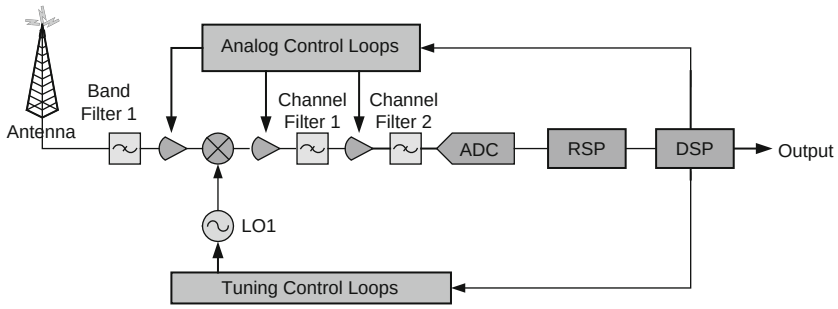


Figure 2.3-7 Single-carrier Rx example.

the RF signal. A common example is found in HF receivers where the desired RF signal may only be a few MHz. In these cases, they are frequently mixed up to IF frequencies of 10.7 MHz, 21.4 MHz, 45 MHz, or higher IF frequencies because of the availability or performance of the required component.) Each successive stage also takes advantage of filtering that is distributed throughout the chain to eliminate undesired images as well as other undesired signals that may have survived the mix down process. The filtering should also be appropriate for the application. A traditional single-carrier receiver would generally apply channel filtering through the mixer stages to help control the IP3 requirements of each stage. Analog channel filtering is not possible in the case of a multi-carrier receiver where the channel bandwidths are not known in advance. Therefore, the mixing process must preserve the entire spectrum of interest. Likewise, our single-carrier SDR application must also preserve the maximum possible spectrum in case the SDR requirements need the full spectrum. In this case, it is probable that our single-carrier example may be processing many carriers, even if only one is of interest. As with the LNA, it would be desirable for the mixer in an SDR to have an adjustable bias. As with the LNA, this bias could be used to properly set the conversion gain and IP3 of the device to correspond to the desired signal conditions.

Some receiver architectures utilize a quadrature demodulator in addition to, or instead of, a mixer. The purpose of the demodulator is to separate the I and Q components. Once they have been separated, the I and Q paths must maintain separate signal conditioning. In the digital domain this is not a problem; in the analog domain, however, the signal paths must be perfectly matched or I/Q imbalances will be introduced, potentially limiting the suitability of the system. Many SDR receivers avoid this problem by utilizing "real" sampling (as opposed to complex sampling), as shown in the single-carrier example, and using a digital quadrature demodulator in the digital preprocessor that will provide perfect quadrature.

The local oscillator (LO) is used to generate the proper IF when mixed with the incoming RF signal. Generally, a LO is variable in frequency and easily programmable via software control using PLL or DDS techniques. There are cases where the LO may not require frequency hopping. One such example is for receiving multiple carriers within a fixed band. In this case, the LO is fixed and the entire band is block converted to the desired IF. It often may be desirable to change the LO drive level to optimize spurious performance under a variety of signal conditions.

Quite often the IF amplifier is in the form of an AGC. The goal of the AGC is to use the maximum gain possible without overdriving the remainder of the signal chain. Sometimes the AGC is controlled from an analog control loop. However, a digital control loop can also be used to implement difficult control loops not possible using analog feedback. In multi-carrier applications, use of an AGC may at best be difficult. If insufficient dynamic range is available in the receiver (determined largely by the ADC), reduction in gain from a strong signal may cause weaker signals to be lost in the noise floor of the receiver. In applications such as this, a digital control loop for the gain is ideal. The control loop can be used as normal as long as no signals are at risk to being lost. However, if a weak signal is detected in the presence of a very strong signal, the decision could be made to allow a limited amount of clipping rather than reduce the gain and risk total loss of the weak signal. Conditional situations like this are much easier to control with a digital control loop than with an analog loop, allowing much greater control of total conversion gain of the receiver.

The ADC is used to convert the IF signal or signals into digital format for processing. Quite often the ADC is the bottleneck, and selection of the ADC is often a driving factor that determines the architecture of the SDR [1, 9, 10]. Oftentimes, the designer is forced to select the best available ADC, realizing that under many conditions the ADC may be overspecified. Still other times, air interface standards may not be directed towards multi-carrier receivers and require much better ADCs than are required when deployed in the field, simply because of the test methodology specified by the standard. For the ADC it may be desirable to change the sample rate, input range, and potentially the active bandwidth.

The digital preprocessor can take many forms. For very high sample and data rates, this is usually implemented as either an FPGA or ASIC. These circuits by nature are quite flexible in their functions and range of parameters. An FPGA can of course be programmed for any function desired. Typically, an FPGA would be programmed to perform the quadrature demodulation and tuning, channel filtering, and data rate reduction. Other functions such as RF power measurement and channel linearization are possible. All of these elements are easily generated using a variety of digital techniques and are readily programmed by loading a variety of coefficients to the FPGA. By doing this, a single-chip configuration can be used to generate a digital preprocessor capable of tuning the entire range of the ADC Nyquist band and filtering a signal with bandwidths from a few kHz to several MHz. When multiple channels are required, the design can be repeated to fill the FPGA. If a lower cost option is required, a variety of ASICs are available that perform these functions. They are often referred to as channelizers, receive signal processors (RSPs), or DDCs.

The final element in the SDR is the DSP. Since this is general purpose DSP, it can be programmed for any

required processing task. Typical tasks include equalization, detection, rake receiver functions, and even network interfacing, to name a few. Because they are fully programmable, they can be used for just about any signal processing task as well as controlling all of the features in the other elements of the block diagram. As DSP processing capabilities increase, DSPs may well take over many of the functions within the digital preprocessors.

Transmit

Transmit functions are also based on some form of super-heterodyne or direct conversion. Figures 2.3-8 and 2.3-9 illustrate these two options. The multi-carrier option is best suited to single- and multi-carrier applications, while the direct conversion offers an excellent, low-cost solution for single-carrier applications. As integration technology improves, multi-carrier direct conversion may become a possibility; however, such a transmit configuration requires sideband suppression that is about 15 dB better than the spurious requirements to prevent images on one side of the center frequency from overpowering a potentially weak carrier on the other.

In either application, a DSP or baseband ASIC is used to generate the modulated baseband data. This data is fed either directly to a pair of baseband DACs (I and Q) for direct RF modulation or to a digital processor responsible for digitally translating them to a suitable digital IF. Depending on the application, the DSP alone or in conjunction with a digital processor can be used to digitally pre-distort the baseband data in such a manner that distortion products generated later in the signal chain will be cancelled.

If an IF stage is employed, the baseband data generated by the DSP must be up-converted either digitally with an FPGA or ASIC (also known as TSPs or DUCs) or alternately with a traditional mixer or modulator to the desired IF. This traditional technique is being replaced by digital means because of the added flexibility offered through digital logic and the availability of good, cost-effective digital-to-analog converters. As with the related receive function, the purpose of this device is to shape the bandwidth of the desired channel and then up-convert by digital means to the desired IF frequency. If multiple channels are required, they can be synthesized on one chip. After translation, each of the channels can be summed together and interpolated to the desired data rate and then sent to a DAC. If desired, digital pre-distortion can be added in conjunction with the DSP to correct for distortion later in the signal chain.

Either a mixer or a modulator is used for frequency translation to the final RF frequency. If direct RF modulation employed, an RF modulator will be used. If an IF is used (either directly from a DAC or a traditional IF up-conversion), a mixer will be used to translate to the final RF frequency. As with the receive mixer/demodulator, it may be desirable to change the bias levels or the drive level of the data or LO levels to optimize distortion.

As with the receive LO, the transmit LO is variable in frequency and easily programmable via software control using PLL or DDS techniques. Here too, it may be desirable to change the LO drive level to optimize spurious performance under a variety of signal conditions. As with the single-band operation of the receiver, there may be cases where a fixed LO is required. Such an example would be for operation within a single band where tuning is accomplished within the transmit signal processor (TSP), DUC, or FPGA.

As with the receive path, the data converter or DAC is often the bottleneck. However, since dynamic range requirements for the transmit signal path are much lower (typically 25 to 45 dB) than the receive path, component selection is not quite as difficult. Many DACs are available that facilitate a wide range of adjustments, including gain and offset correction so that I/Q imbalances in the transmit signal chain can be minimized. Other desired features include data rate interpolation and I/Q phase correction.

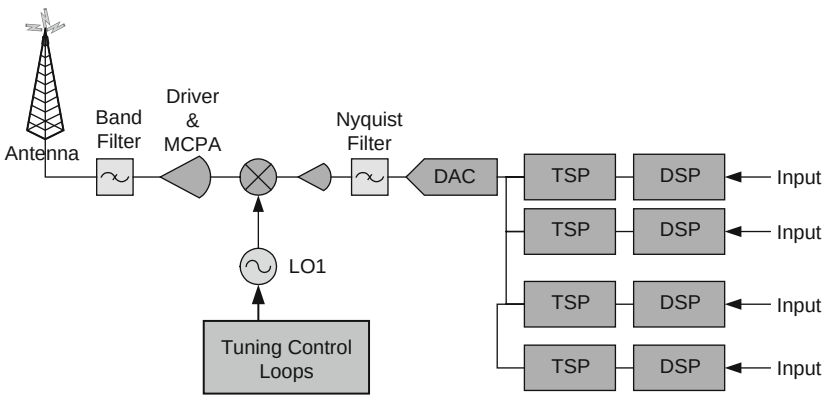


Figure 2.3-8 Multi-channel transmit with single up-convert super-heterodyne.

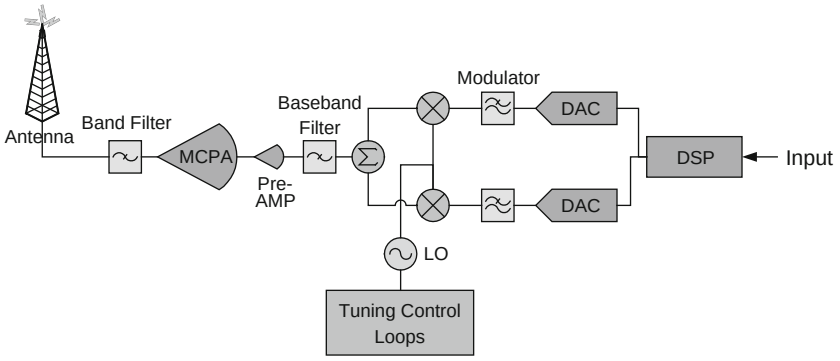


Figure 2.3-9 Single-carrier direct-conversion transmit.

Finally, power gain is achieved through a pre-amp and PA. Aside from the fact that these devices must operate across a wide range of frequencies, it is desirable to adjust the RF output power. There could be regulatory issues that require some frequencies to be transmitted at lower power than others. While the PA gain is usually fixed, the pre-amp may be in the form of a VGA.

Implementation issues

The reality is that without improvements in semiconductor technology through the late 1990s, SDR as outlined here would still be a concept and not a reality. Although the evolution of DSP technology has certainly been key to SDR, it is not the only technology that has had to "grow up" in order to support this development. Because the level planning is different in these systems, most of the analog components are stressed to a higher degree, and better performance is required than that found in traditional architectures.

Analog front end

Quick conversion

In order to take advantage of digital processing, an SDR seeks to convert from the RF domain to digital domain as soon as possible. This is true both for receive and transmit. In doing this, as much of the processing can be done digitally as possible.

Advantages

If most of the processing is done digitally, then reconfiguration can be quite simple. Filter coefficients can be changed, and different software can be run or even, in the case of FPGAs, completely reconfigured for the required format. In the analog domain, space and resources limit the reconfiguration options available.

In the analog domain, only a small number of modulation and demodulation schemes are possible. However, in the digital domain, the possibilities are limitless if the functions are configurable through software. Even where complex functions are implemented in the analog domain, various errors such as quadrature errors can cause performance issues.

In the digital domain, quadrature and other functions can be exactly generated. Once in the digital domain, the accuracy of the function is limited only by the bit precision of the math used to implement the function. For example, it is always much easier to add a few more bits to a digital multiplier than it is to improve the linearity of an analog mixer.

Finally, because most of the signal chain is digital, performance would be more consistent for each system manufactured, eliminating much of the product variation and yield. Since the performance is more consistent, many of the factory trim and alignment issues would be eliminated, potentially reducing a large part of the manufacturing and test cost.

Higher gain at RF frequencies

These advantages are good for providing a more consistent transceiver at a lower production cost. However, as already mentioned, the level planning for an SDR is sufficiently different from a standard superheterodyne that different strategies are required.

Level planning

In a traditional super-heterodyne transceiver, the conversion gain is distributed throughout the signal chain. Typically, gain in the front end is balanced between high enough for a low NF but not so high as to overdrive the remainder of the signal chain and cause excessive intermodulation distortion. Similarly, as much gain is run after the channel filtering as possible so that interfering signals have already been eliminated. Throughout the signal chain, only enough gain is used to offset the accumulated losses and to prevent those elements from significantly contributing to the overall NF, thereby allowing for IP3 to be carefully balanced against NF.

By contrast, in an SDR many of the intermediate stages have been eliminated because sampling occurs as close to the antenna as possible. Since most SDR applications are also wideband, there are no channel filters, thereby allowing many of the neighboring signals to also pass the signal chain. Because all of the conversion gain must occur in the presence of many signals, intermodulation performance is inherently more important.

In a transmitter with multistage up-conversion, the issues are very similar to the super-heterodyne receiver above. As with receivers, both noise and intermodulation are very important specifications. However, the active dynamic range of most transmit requirements are only on the order of 60 to 70 dB, whereas most receivers require 100 dB or more of dynamic range. The real difficulty of the transmit signal path is maintaining the noise and linearity performance in the RF power amplifiers when the power level reaches several hundred or even several thousand watts. The discussion of the PA is beyond the scope of this chapter.

Fixed versus variable gain

In a traditional receiver, the total conversion gain is quite often variable. This reduces the required dynamic range of the circuitry following the AGC, reducing the required linearity requirement of those components. Also, the AGC action allows for optimum signal levels over a wider range of input signal conditions. While an AGC is still quite useful for SDR, there are certain restrictions on their use.

Tradeoffs versus MC and SC

In a single-carrier receiver there are two main issues with setting the gain. The first issue is that it is desirable not to overdrive the front end when a strong signal of any frequency is present. After this issue is accounted for, the conversion gain can be increased or decreased as necessary to achieve the sensitivity required. Although it is possible that a nearby signal will pass the front-end filters of the receiver and cause a reduction of the gain and subsequent loss of the desired signal, this is typically managed through the use of "band select" or RF tracking filters that filter all but the desired signals. However, SDR and multi-carrier receivers typically have a "wider" front-end bandwidth and therefore allow many more signals to pass through the full analog signal chain. As a result, it is much easier for a strong signal at one frequency to desensitize the desired signal at another frequency. Since the receiver has a limited noise floor (thermal and other noise sources), the gain can only be reduced to the point that the weakest signal retains the minimum SNR required for detections. Since the design has already been configured for multiple carriers, it is likely that the gain has been reduced to a minimum so that the largest expected signal will not overdrive the signal chain. Because the gain is limited, the noise floor of the receiver becomes limited by that of the data converter.

Tradeoffs versus converter resolution

In a traditional receiver, if the signal level was not large enough to be adequately detected by the ADC, then additional gain is used to boost the level above the ADC noise floor using an AGC topology. However, it is just as valid to lower the noise floor of the converter. There are several ways to do this, as discussed in what follows. The easiest thing to do is just specify a converter with more bits. Unfortunately, the more bits a converter has, the more expensive it is and the more power it dissipates. Therefore, balancing the conversion gain and converter resolution is a very important task: too much gain and the ADC is overdriven; too little gain and the ADC directly sets the noise floor, which is an undesirable situation [11]. Ideally, the conversion gain of the receiver should place non-ADC noise 10 dB above the ADC noise. Therefore, given an ADC converter noise floor, an ideal minimum gain can be determined that prevents the ADC from dominating overall noise performance.

The equation for maximum conversion gain in dB is

$$G_{\max} = P_{\text{ADC_Fullscale}} - P_{\max_\text{signal}} \tag{2.3.1}$$

The equation for minimum conversion gain in dB, on the other hand, is

$$G_{\min} = \text{NSD}_{\text{ADC}} + 10 - 10\log\left(\frac{k \times T \times 1\ \text{Hz}}{0.001}\right) - \text{NF}_{\text{AFE}} \tag{2.3.2}$$

These equations outline the desired maximum and minimum conversion gain. To achieve gains beyond these bounds, an AGC can be carefully used. Even with an AGC, the data converter will determine what the instantaneous dynamic range of the transceiver will be based on the difference between the noise floor and the range of the converter. If $G_{\min}$ is greater than $G_{\max}$, either the range of the ADC must be increased or the NSD (noise spectral density) of the ADC must be lowered, indicating that a better converter may be required. Although there are means of increasing the maximum input of the converter or reducing the noise floor of the ADC, it is often easier to specify an ADC with better performance or more bits of resolution.

Higher IP3 requirements

Although neither option comes easily, increasing the fullscale of the converter may have other undesirable consequences. If the input range is increased, then larger signal swings are required to take advantage of this

increased range. This implies that high signal powers are required. Therefore, in order to keep the intermodulation products at the same level, the IP3 specification of the drive circuitry must also be increased to take full advantage of the increased signal range, otherwise, what signal dynamic range is gained will be quickly lost to increasing intermodulation distortion, most notably the 3rd-order products. For IF sampling, even order intermodulation products are generally not a problem because they most often fall away from the signal of interest and are easily filtered from the spectrum of the input to the ADC input.

Signal dynamic range

Signal dynamic range is the difference between the largest and the smallest signal that can be detected. If the receiver is properly designed, the ADC will determine the dynamic range. The fullscale range of the ADC will determine the largest signal as already established by reworking Equation (2.3.1). Likewise, the smallest detectable signal will be set directly or indirectly by the noise floor of the converter. Ideally, the noise from the AFE will dominate the total noise because it has been placed as much as 10 dB above the converter noise floor. If the front-end noise floor is much less than 10 dB, then the contribution to total noise from the ADC will increase and must be included in the overall noise calculation of the receiver.

The largest signal is determined by the fullscale of the ADC and the applied conversion gain. Similarly, the smallest signal to be detected can be calculated by the noise in the channel of interest. The maximum detectable signal before clipping is determined by:

$$\mathrm{XDS} = P_{\mathrm{ADC_Fullscale}} - G \tag{2.3.3}$$

The minimum detectable signal is determined by:

$$\mathrm{MDS} = 10\log\left(\frac{k \times T \times \mathrm{BW}_{\mathrm{signal}}}{0.001}\right) + \mathrm{NF}_{\mathrm{cascaded_total}} + \mathrm{SNR}_{\mathrm{required}} \tag{2.3.4}$$

For example, if the ADC has a fullscale of +4 dBm and the conversion gain is 35 dB, the maximum input power is −31 dBm. Similarly, if the channel of interest is 200 kHz wide, the total NF is 3 dB and the required SNR is 5 dB, then the MDS is −112.8 dBm. Taking the difference between Equations (2.3.3) and (2.3.4) will estimate the dynamic range of the receiver. In this example, the signal dynamic range is found to be 82 dB:

$$\mathrm{DR} = P_{\mathrm{ADC_Fullscale}} - G - 10\log\left(\frac{k \times T \times \mathrm{BW}_{\mathrm{signal}}}{0.001}\right) - \mathrm{NF}_{\mathrm{cascaded_total}} - \mathrm{SNR}_{\mathrm{required}} \tag{2.3.5}$$

There are many factors that will reduce both the MDS and the DR. A key point to remember is that, as shown here, XDS, MDS, and DR are "static" tests and that in reality more than one carrier may share the dynamic range of a multi-carrier receiver. Because of constructive interference, the fullscale power of the converter must be shared between each of the signals, thereby effectively reducing the largest possible input signal. As a guideline, if all signals are at the same level, each time the number of carriers is doubled, the largest any of them can be is reduced by 3 dB of power. For example, if two signals are present, each signal must be 3 dB below the clipping point. If 4 are present, 6 dB and 8 must have 9 dB and so on. Therefore, for applications where many signals are present, the effective dynamic range is limited.

Similarly, there are reasons that the noise floor will increase above that calculated in Equation (2.3.4). One example of this is seen through reciprocal mixing between the phase noise of the LO and a nearby blocking signal, resulting in an increase in the noise floor of the mixer. A similar example is seen in the increase in the converter noise floor associated with the "reciprocal mixing" between the same blocker and the aperture jitter of the ADC. Fortunately, if the converter noise floor is adequately below that of the AFE, variations of several dB in the noise of the ADC will have only limited effects in overall performance. If, however, the ADC was designed to dominate overall noise or guarding of much less than 10 dB was used between the two, the effects on the overall receiver performance should be revisited to determine the effects of ADC noise versus a variety of signal conditions [11].

IF Selection/tradeoffs

Selecting an IF frequency for a traditional single-carrier receiver can be challenging enough. However, in a multi-carrier receiver, traditional issues such as determining the level of the high-order intermodulation products of all signals is complicated by the fact that now entire bands are being translated from one frequency to another. This problem is further complicated by aliasing within the ADC. In a typical single-carrier receiver, the IF signal into the ADC is chosen such that any aliased harmonics of the input signal fall away from the input signal. This is important at low signal levels because when an ADC is stimulated by very low-level inputs, it is possible that spurious responses generated within the ADC can be larger than the desired input. Thus if the harmonics are designed to fall away from the desired input, this problem can be averted. However, with a multi-carrier receiver, the harmonics can cover a very wide band of frequencies. Generally, even-order harmonics are not much of a problem. At very low signal

levels, the key harmonics are the third and fifth harmonics of the input spectrum. Since the input may be a wide band, the third and fifth harmonics are 3 and 5 times as wide, respectively. Given this, it becomes difficult to try and place these signals in a part of the spectrum where they will cause no problems. In cases where careful signal planning is not possible, dithering techniques provide relief to many of these problems [12]. Figure 2.3-10 shows how it is possible to place some harmonics out of band if sufficiently oversampled. In this example, the second and third harmonics are placed in such a manner that the aliased components fall away from the desired fundamentals.

Transition band requirements

Since all of the channel selection in an SDR is done in the digital domain, the analog filter requirements are different. Their purpose is primarily to prevent the overlap of images either in the mixing process or the sampling process. In mixer stages, care must be taken to suppress the undesired mixer images. In the ADC, signals both above and below the desired band may be sampled and aliased into the usable spectrum of the ADC. For either mixer images or aliased signals within the ADC, these signals must be filtered so that they are below the minimum detectable signal. If they are not, then it is possible that they will overpower the desired signal. In the case of the ADC, this required rejection must be achieved before the aliasing becomes critical. As shown in Figure 2.3-11, the full rejection of the undesired signals must be achieved before the spectrum is folded upon itself.

Data conversion

For both transmit and receive, the data converters are usually key components in the signal plan. The key elements for both are the dynamic range, which is bound on one end by the noise floor of the converter and the maximum input or output range on the other. Other equally important issues include distortion, both harmonic and intermodulation. Although related, they have somewhat different effects in the limitation of performance.

There are many different converter topologies and each has their benefits and limitations [13, 14]. While there is no set architecture that provides better performance for SDR, the selection is best made based on the performance requirements for the application and then a study of the available data sheets of the potential converters. Since data sheets cannot fully represent the actual performance of a data converter, it is always best to take a look at them on the bench in an environment similar to that of the end product.

Because data converters are somewhere between the digital and analog domains, they are often poorly understood by both the analog designer and the digital designer. For this reason, their effects on transceiver design are often overestimated, underestimated, or both in different areas. The next few topics will sort through many of the issues to help determine exactly how converter performance determines performance.

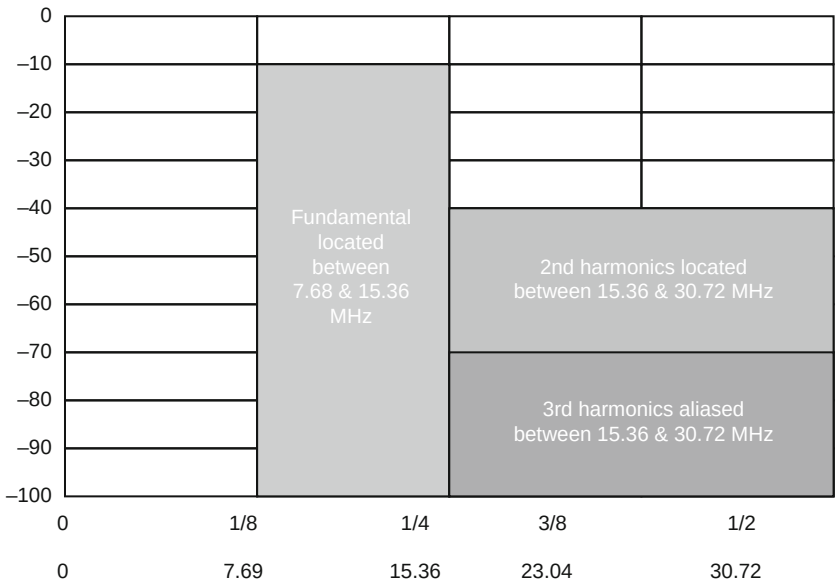


Figure 2.3-10 Aliased wideband signal and harmonics.

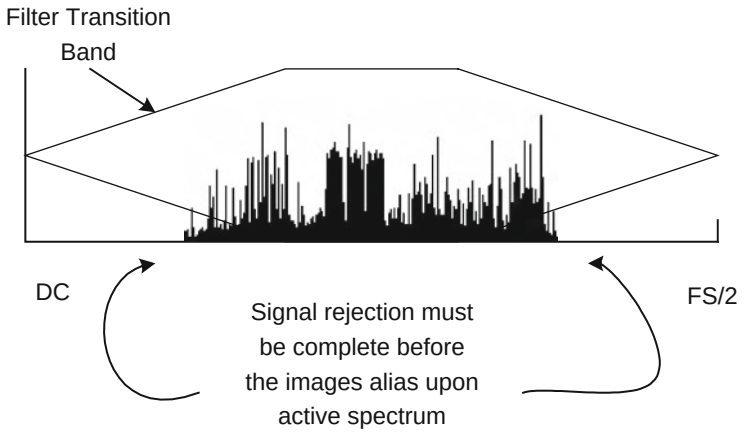


Figure 2.3-11 Aliased transition bands.

General converter requirements

Bits and noise and a little signal

The number of converter bits is the most visible specification. From a mathematical point of view the number of bits the converter contains limits performance physically. An ideal converter will exhibit an SNR that is determined by this equation:

$$\text{SNR} = 6.02N + 1.8 \tag{2.3.6}$$

In reality, there are many other issues that determine the performance of the converter, including clock jitter and thermal noise. Equation (2.3.6) provides the noise due to ideal quantization and does not take into account any of the other sources of noise in a converter. A modification of this equation provides a more insightful measure of converter performance. Equation (2.3.7) for modified converter SNR takes into account clock jitter, thermal noise, and the effects of non-ideal quantization, which are the major limitations in converter performance.

$$\text{SNR} = -20\log\left[(2\pi F_{\text{analog}} t_{j_{\text{rms}}})^2 + \left(\frac{1+\varepsilon}{2^N}\right)^2 + \left(\frac{v_{\text{noise}_{\text{rms}}}}{2^N}\right)^2\right]^{\frac{1}{2}} \tag{2.3.7}$$

Although this equation is a little more complicated, it does take into account many of the important factors in converter performance. In Equation (2.3.7), F_{analog} is the analog IF frequency, $t_{j_{\text{rms}}}$ is the aperture uncertainty, ε is the average DNL of the converter, $v_{\text{noise}_{\text{rms}}}$ is the thermal noise of the converter, and N is the number of bits.

Many data sheets will include information on the effective number of bits that a converter possesses. This number is usually determined by solving Equation (2.3.6) for N, the number of bits. While an effective bits measurement is a convenient comparison tool, it has limited usefulness in radio design. SNR is a better tool because of its direct link to noise. A measurement that is better still is that of NSD for the converter. NSD provides the amount of noise energy in a 1 Hz bandwidth for a converter. This number is not usually specified in the data sheet because it is dependent on the actual sample rate used and the input termination condition. To determine the NSD of a data converter,

$$\text{NSD} = P_{\text{ADC_Fullscale}} - \text{SNR}_{\text{ADC_Fullscale}} - 10\log\left(\frac{\text{Sample_Rate}}{2}\right) \tag{2.3.8}$$

Once the NSD has been determined, it can be used either to confirm if the converter meets the noise floor requirements or to determine the minimum gain required from the front end of the transceiver design.

As an example, an ADC is selected with an SNR of 70 dB at the selected input frequency. With the required input termination, a fullscale of +4dBm is achieved. The configuration requires a sample rate of 61.44 MSPS. Using Equation (2.3.8), the NSD is −140.9 dBm/Hz.

As outlined above, if it is desirable that the converter noise not limit the performance of the transceiver, then the front-end noise needs to be approximately 10 dB higher than that of the converter. Therefore, to safely use this converter, noise generated by the front end must be about −131 dBm/Hz. Modifying Equation (2.3.2) because the NF of the AFE is not yet known, gives Equation (2.3.9), and the gain and noise requirements. With this equation, the combined front-end gain and noise figure

must be 43 dB to ensure that the ADC does not dominate performance.

$$G_{min} + NF_{AFE} = NF_{ADC} + 10 - 10\log\left(\frac{k \times T \times 1\ \text{Hz}}{0.001}\right) \quad (2.3.9)$$

Noise figure for an ADC

If it is not possible to design the system such that the converter noise is significantly below the remainder of the system, then the noise must be included in the calculation. This can be accomplished by using Equation (2.3.9) to determine the noise from the ADC, or the NF of the converter can be calculated and included in the cascaded NF of the signal chain analysis with the other linear devices. While an ADC is not a power device, the NF can be estimated and should only be considered valid for the set of operating conditions specified. Therefore, if the conditions are changed, then the NF will change appropriately. The equation for the equivalent ADC noise figure is:

$$NF_{ADC} = P_{ADC_Fullscale} - SNR_{ADC_Fullscale} - 10\log\left(\frac{\text{Sample_Rate}}{2}\right) - 10\log\left(\frac{k \times T \times 1\ \text{Hz}}{0.001}\right) \quad (2.3.10)$$

If it is determined that the conversion gain required to offset the converter noise is large enough that the converter is overdriven, this is an indication that a better converter is required.

Channel noise in a receiver

Once the total receiver noise level has been determined, sensitivity of the receiver can be found. If conversion gain is known, then sensitivity with respect to the antenna can be found; otherwise, it will be with respect to the ADC input. In the typical SDR signal chain, a digital tuner or channelizer will follow the ADC. In this block, the desired signal is tuned and all other signals in the Nyquist band are filtered from the spectrum. Typically the data rate is also reduced to a speed that is suitable to the data rate of the modulation being carried. If a quiet channel is selected, all that should be on the output of the channelizer is the noise from the AFE plus the ADC. Since the NSD has already been established in a prior section, the total noise in the channel can be determined by integrating this over the channel bandwidth. In log math, the equation for receiver channel noise is very simple:

$$N_{channel} = 10\log\left(\frac{k \times T \times 1\ \text{Hz}}{0.001}\right) + G_{min} + NF_{total} + 10\log\left(\frac{BW_{channel}}{1\text{Hz}}\right) \quad (2.3.11)$$

With the total integrated channel noise and the required SNR for the modulation standard, the reference sensitivity can be determined. Keep in mind that the required SNR may be positive or negative depending on the amount of digital gain provided by detection algorithm. As an example, GSM requires about 5 dB SNR while IS-95 requires an SNR of about −16 dB.

If the AFE generates a noise density of −131 dBm/Hz (the ADC is 10 dB below this and not a contributing factor) and the channel bandwidth is 200 kHz, then the total channel noise is −78 dBm/200 kHz. If the required signal level is 5 dB above this, the smallest signal that can be detected as presented to the ADC will be −73 dBm. If the conversion gain of our signal chain is known, then the sensitivity at the antenna can be calculated. In order to achieve the noise of −131 dBm/Hz, a gain plus NF of 43 is required. At this point the NF is not known but may be estimated based on available technology. A good typical NF would be about 3 to 4 dB. This would place the conversion gain at 40 dB. Therefore, if the −73 dBm signal is referred back to the antenna, it will be 40 dB smaller or −113 dBm, a very good sensitivity for a channel 200 kHz wide.

Digital processing

Parts

There are two categories of digital parts. Both could be called DSPs. The traditional DSP is a computational unit that consists of program and data memory. A program is executed from the program space operating on data from I/O ports and data stored in the data memory. This type of DSP is the most common; however, this type of DSP is limited in the data throughput. While great advances in parallel computing and core speeds have increased the rate at which real time data can be processed, general purpose DSPs can only process limited amounts of data.

Fixed function DSP

To augment the processing power of a general purpose DSP, fixed function DSPs are designed to process very large amounts of data very fast and efficiently. While a general purpose DSP can be infinitely reprogrammed, the signal flow within a fixed function DSP must be restricted to a single architecture. Programming is also limited to configuration registers and memory coefficients. However, since most radios are based on some form of super-heterodyne architecture, this is not such a limitation due to the high degree of similarity between different designs. Therefore, the fixed function DSPs can be designed to represent a very large class of receiver or transmitter designs. These fixed function DSPs are often

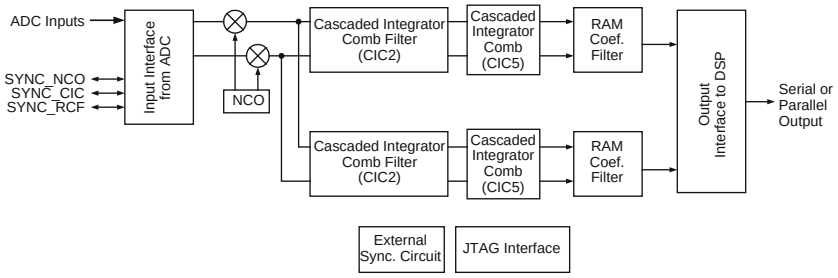


Figure 2.3-12 Fixed function receive signal processor (RSP).

implemented either in FPGAs or ASICs. Processors designed for the receive function are called receive signal processors (RSPs), as shown in Figure 2.3-12, and transmit functions are called transmit signal processors (TSPs), as shown in Figure 2.3-13. In general, both TSPs and RSPs contain exactly the same elements, only the order is reversed. For either device, there are three key subfunctions found in these devices.

The first function is the frequency translation. In the analog domain, the frequency is translated with a mixer or modulator/demodulator. This function is used to mix two inputs together in such a way that the sum and difference frequencies are generated on the output. In the digital domain, this is represented by a multiplication. If the function is a mix, then a single "real" multiply is performed, but most often, the multiply is a complex multiply used to generate quadrature data and thus separate positive and negative frequencies. In an IF sampling receiver, a real digital IF is applied to one of the complex inputs of the multiply. The other input is the output from a complex numerically controlled oscillator (NCO). The NCO is tuned to the desired frequency such that the result is a complex signal at DC and at the sum frequency.

Following the NCO and complex mixer (demodulator) is a low-pass filter. This filter serves two purposes. The first purpose is to remove the undesired noise, signals and spurious. By doing so, all of the wideband noise on the output of the ADC is removed except that which lies within the passband of the filter, giving rise to what is often referred to as processing gain. Second, the filter shapes the passband characteristics. Quite often, the passband must be a matched filter or otherwise shape the characteristics of the incoming spectrum. This is easily accomplished with a digital filter, a task that is often difficult with analog channel filters. In fact, since these filters are digital, they can implement any filter that can be realized using FIR or IIR techniques. This filter is determined by Equation (2.3.12):

$$G_{\text{Noise_Processing}} = 10\log\left(\frac{\text{Sample_Rate}/2}{\text{BW}_{\text{channel}}}\right) \tag{2.3.12}$$

Following channel filtering, the bandwidth will be relatively small compared to the data rate because of the high oversampling rate in the ADC. Therefore, it is advantageous to reduce the data rates. This has several benefits. First, with the reduced data rate, the computational burden on the general purpose DSP is reduced. Second, in CMOS technology, lower data rates result in lower power. Therefore, following the channel filters, data decimation is performed. The decimation must be consistent with the Nyquist, but significantly reduces the computation by the general purpose DSP that follows the RSP.

In the transmit direction, the data flow is reversed. First the data is filtered and then interpolated to the desired data rate. Then, the data is translated to the proper frequency using a modulator and complex NCO.

For a typical RSP/TSP channel, the computational load may be as high as 1.5 Giga-operations per second. If multiple channels are required, then the process scales

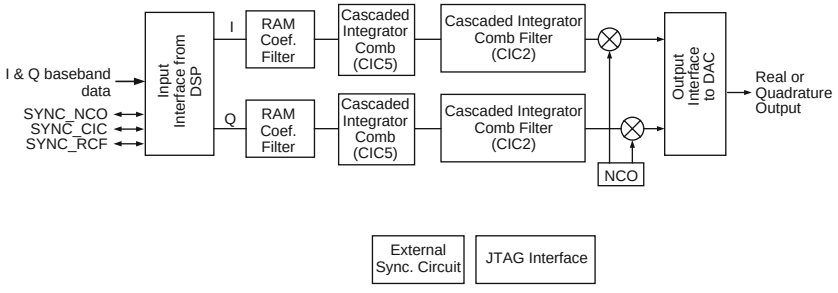


Figure 2.3-13 Fixed function transmit signal processor (TSP).

linearly. At the present time, this load exceeds the capabilities of a general purpose DSP; however, as DSP; technology improves, it may be possible to take on some or all of the processing in the future.

General purpose DSP

General purposes DSPs like microprocessors are designed to execute a software program. The software for a DSP is developed in the same manner as that for a microprocessor using program languages such as "C" and assembly. However, DSPs are designed specifically to execute code as fast as possible. In fact, DSPs are usually designed to execute programs or tasks in real time. Since the DSP may be processing realtime data such as voice or video, program execution is required to keep up with incoming data, otherwise throughput will be sluggish, intermittent, or simply come to a halt as the DSP struggles to keep up with incoming data. A block diagram of a typical DSP is shown in Figure 2.3-14.

To prevent this from happening, DSPs are especially designed to improve data throughput, taking advantage of a number of techniques. Often, one vendor will focus on one technique and refine that while a different vendor will focus on a different optimization technique. Both result in faster throughput, but with slightly different advantages. Some of these techniques are listed here:

- Program and data look-ahead caching
- Multiple address generation
- Separate program and data memory
- Multiple arithmetic logic units
- Multiply accelerators
- Shifters
- Hardware accelerators
- On-chip memory
- Floating/fixed point optimization

In an SDR, the general purpose DSP is generally tasked to perform the Nyquist rate processing; that is, the signal processing required at data rates that supports the Nyquist rate of the raw data. In our SDR application that may support a channel as wide as 10 MHz, the actual data rate may be as high as 20 MHz. While not all applications may require this much processing, some applications may. The actual processing requirements will depend on the application and functions instantiated. As with other components, if a wide range of processing is expected, the design has to consider the maximum requirement, even in the case where excess processing capability may exist in some operating modes:

- Envelope detection (AM)
- Phase/frequency detection (PM/FM)
- Phase/frequency correction
- Equalization of a TDM burst
- Spread/de-spread a CDMA signal
- Voice coding/decoding

These are just a few of the functions that are typically performed. Since the DSP is programmable, any function that can be coded can be executed. Additionally,

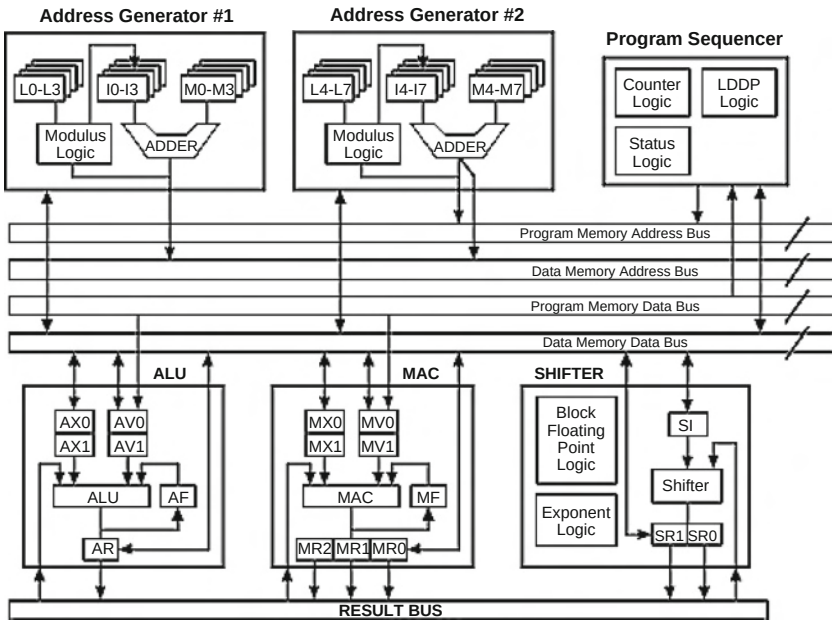


Figure 2.3-14 General purpose digital signal processor (DSP).

since the code is software, it can be upgraded or changed at any time to further support the SDR.

Case study: A close look at a CDMA2000 and UMTS SDR receiver

Now that many of the facets of SDR have been discussed, the final section will cover an example of a multi-carrier SDR receiver. Although this is not a full analysis, it will cover many of the issues that surround the design and development of an SDR that are not covered in a typical receiver design. As with any design, the first place to start is with the specifications. The following table summarizes a few of the critical specifications for both CDMA2000 and UMTS (WCDMA).

	CDMA2000	UMTS
Reference sensitivity	−117 dBm	−121 dBm
Bandwidth	1.25 MHz	5 MHz
Chip rate	1.2288 MCPS	3.84 MCPS
Signal spectral density	−177.9 dBm/Hz	−186.8 dBm/Hz
Sample rate	61.44 MSPS 50 × oversample	61.44 MSPS 16 × oversample
Despread and coding gain	21 dB	25 dB
Narrowband blocker	−30 dBm	Na
CDMA blocker	−52 dBm	−40 dBm
Two-tone blocking	−45 dBm, 2 tone CW	−48 dBm, 1 CW and 1 CDMA

The goal of this exercise will be to design a multi-carrier, multimode, single-band receiver RF through baseband that is capable of processing both of these standards either independently or at the same time. Such a design would be useful for manufacturers of 3G base station equipment where it is desirable to have a single piece of hardware that is capable of processing both standards, thereby eliminating duplicated design efforts.

One of the most direct ways of accomplishing this is to compare the two specifications and determine which will limit performance. One of the first issues will be to determine the largest signal that requires processing. The CDMA2000 standard calls for a −30 dBm narrowband signal, whereas UMTS does not address narrowband blockers. However, it does require that a −40 dBm CDMA signal be correctly processed. While narrowband signals can often be considered to have little envelope, a CDMA signal has between 10 and 12 dB of peak to rms on the envelope. A CDMA signal of −40 dBm thus actually peaks very close to −30 dBm. Therefore, both standards require about the same peak signal capacity.

Since we know that we will need to digitize the signals, initial ADC characteristics may be established. Later in the analysis, the specification can be validated to determine if the assumptions were correct. Since high-performance data converters are expensive, it is desirable to use the lowest performance possible that allows the specifications to be met. In typical converter specifications [19], the fullscale input range is 2 volts peak to peak differential. If this input is terminated with 200 ohms differentially, the rms power to drive the converter to fullscale will be +4 dBm. Similarly, the converter SNR is 75 dB and the SFDR, both single and two-tone, is −95 dBFS. This performance is maintained out to analog frequencies of 100 MHz providing flat performance.

From this information an initial estimate of the conversion gain required can be determined using Equation (2.3.1):

$$G_{\max} = P_{\text{ADC_Fullscale}} - P_{\text{max_signal}} = +4 - (-30) = 34 \text{ dB}$$

In order that the receiver not be overdriven, the conversion gain will be limited to 30 dB. For the moment, assuming that the noise figure of the front end, less the ADC, will be 3 dB, the thermal noise delivered to the ADC can be determined. At room temperature, the thermal noise can be calculated to be

$$\begin{aligned}\text{NSD}_{\text{AFE}} &= 10\log\left(\frac{k \times T \times 1\text{Hz}}{0.001}\right) + G + \text{NF}_{\text{AFE}} \\ &= -174 \text{ dBm/Hz} + 30 + 3 \\ &= -141 \text{ dBm/Hz}\end{aligned}$$

Based on this information, the NSD of the ADC can be determined using Equation (2.3.8):

$$\begin{aligned}\text{NSD} &= P_{\text{ADC_Fullscale}} - \text{SNR}_{\text{ADC_Fullscale}} - 10\log\left(\frac{\text{Sample_Rate}}{2}\right) = 4 \text{ dBm} - 75 \text{ dB} \\ &\quad - 10\log\left(\frac{61.44 \text{ MHz}}{2}\right) = -145.9 \text{ dBm/Hz}\end{aligned}$$

Since the NSD of the ADC is less than 10 dB better than the NSD of the AFE, the noise contributed from the ADC must be included in the overall noise analysis. Therefore, using Equation (2.3.10) will provide the equivalent NF of the ADC for the configuration used here.

$$\begin{aligned}
NF_{ADC} &= P_{ADC_Fullscale} - SNR_{ADC_Fullscale} \\
&\quad - 10\log\left(\frac{Sample_Rate}{2}\right) \\
&\quad - 10\log\left(\frac{k \times T \times 1\ \text{Hz}}{0.001}\right) \\
&= 4\ \text{dBm} - 75\ \text{dB} - 10\log\left(\frac{61.44\ \text{MSPS}}{2}\right) \\
&\quad - 10\log\left(\frac{1.38 \times 10^{-23} \times 300 \times 1\text{Hz}}{0.001}\right) \\
&= 28\ \text{dB}
\end{aligned}$$

Based on this information and commercially available components, the level planning in Figure 2.3-15 can now be generated. This design features double conversion in the analog domain to allow for more efficient processing of images and out-of-band blockers. Additionally, dual down conversion offers the possibilities of producing an IF frequency in the range that the ADC can faithfully digitize.

A traditional numerical analysis of this signal chain shown in Figure 2.3-15 provides the following results:

Total NF	4.13 dB
Gain	30 dB
Input IP3	−8.7 dBm
Output IP3	+21.3 dBm

Given this signal chain, the SNR can now be determined for the reference sensitivity on p. 170. Using these updated terms in the equation above for the overall NSD, the total noise can now be determined.

$$\begin{aligned}
NSD_{AFE} &= 10\log\left(\frac{k \times T \times 1\text{Hz}}{0.001}\right) + G + NF_{AFE} \\
&= -174\ \text{dBm/Hz} + 30 + 4.13 \\
&= -139.87\ \text{dBm/Hz}
\end{aligned}$$

If this energy is integrated over the chip rate for each of the standards, the total noise in the channel for each standard will correspond to the data in the following table.

	CDMA2000	UMTS
Noise at antenna	−113.1 dBm	−108.2 dBm
Noise after ADC	−78.9 dBm	−74.0 dBm
Signal energy after ADC	−87 dBm	−91 dBm
SNR after ADC	−8.1 dB	−16.9 dB
Required SNR (approx.)	−16 dB	−19 dB

In both cases, adequate SNR is maintained, resulting in adequate sensitivity. Additionally, the excess SNR can be used to increase the sensitivity of the receiver beyond that for which the specification calls.

In addition to sensitivity, the spurious performance of the signal chain must be analyzed. The analysis of spurious performance is a little bit more difficult but can nonetheless be analyzed. In analyzing the CDMA2000 specifications, there are two specifications to review. The key specifications are the two-tone blocking and single-tone blocking tests.

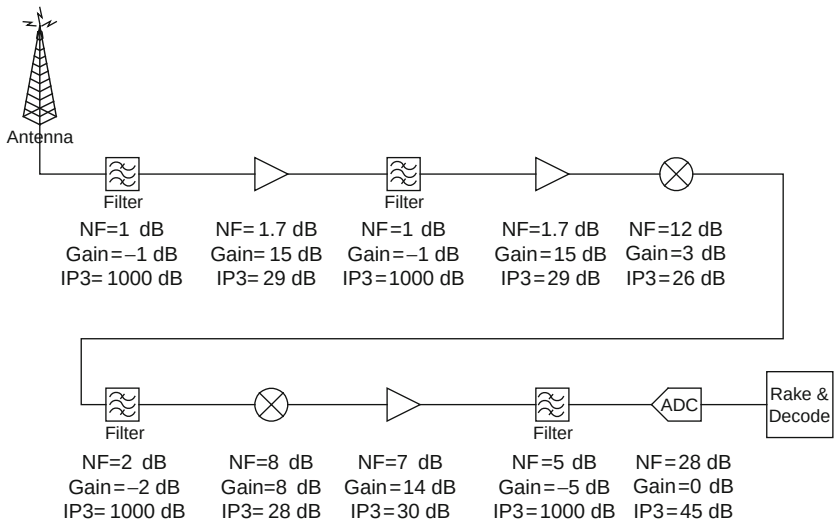


Figure 2.3-15 SDR receive signal chain.

Two-tone blocking requires that the receiver tolerate two CW carriers at −45 dBm. Since this is an IF sampling application, the even-order term (difference) falls near DC and is filtered. The odd-order products are the most critical, especially the third-order products that fall in-band. In Figure 2.3-16, a third-order term (2f1-f2) is shown to fall near the channel center of the CDMA channel. Since it falls near the channel center, it cannot be allowed to disrupt the desired CDMA carrier. The goal is to determine how large the intermodulation product can be such that disruption does not occur. Fortunately, because the desired carrier is a CDMA signal, it will pass through a despreading circuit. This has a side benefit—the decorrelation of the undesired spurious term. After decorrelating, the CW signal will resemble white noise, as shown in the right half of the drawing, while the desired CDMA signal will be rendered as a narrowband and easily filtered and processed. Since the spurious signal becomes pseudorandom noise, it adds to the effective thermal noise at a density of −174 dBm/Hz (kT noise). Furthermore, the mobile power is allowed to increase by 3 dB during this test indicating that the noise generated by the spurious signal is allowed to equal that of the thermal noise. If it is assumed that the spurious products are generated in the ADC, then the noise figure may be added to the thermal noise before determining how large the spurious signal can be. Reflecting all the spurious signal to the antenna, the effective thermal noise including the NF of the entire signal chain produces an NSD of −169.87 dBm/Hz. Integrating this over 1.25 MHz will provide the total energy that may be contained in the spurious signal without adversely effecting performance of the receiver. The total power in 1.25 MHz is −108.9 dBm/Hz. This is the spurious level reflected to the antenna that will not cause blocking in the receiver. Since the receiver is blind as to how the spurious signal is generated, this number is valid for single- or two-tone blocking. Therefore, since the two CW tones were at −45 dBm, the input referred to IP3 is found to be −13 dBm or −63.9 dBc. Likewise the single-tone performance can be calculated with reference to −30 dBm, giving −78.9 dBc. Since the ADC performance is listed at −95 dBc for either single or two tones, no performance limitations should be anticipated.

UMTS is a little different. First, there is no single-tone desensitization specification. Thus the primary specification is the intermodulation testing. This too is different because only one of the tones is CW; the other is a modulated CDMA carrier. Therefore, when 2f1-f2 is generated, it is an image of the CDMA signal shifted by the difference between itself and the CW tone. If this image falls directly on top of a desired CDMA carrier, it is possible that if the chipping sequence is not orthogonal to that of the desired carrier, then the blocking signal could be received instead of the desired signal. More likely, however, is that the unde-sired signal will simply increase the effective thermal noise. As shown in the right half of Figure 2.3-17, as with the CDMA2000 example, the desired signal is correlated and the undesired signal is spread. If the data rate is high enough into the despreading device, the original wideband intermodulation product will become doubly spread as it is convolved with the orthogonal despreading code. However, most often the over-sampling ratio into the despreading function is only 2 or 4, potentially causing much of the doubly spread energy to alias back into the band of interest, thus causing insignificant decrease in the spectral density of the noise. As with CDMA2000, the mobile is allowed to increase its power by 3 dB, indicating that the noise due to the intermodulation product can equal the thermal noise.

Integrating the noise of 3.84 MHz gives a total intermodulation noise of −108.1 dBm. Again allowing for the noise figure of the Rx chain of 4.13 dB allows this noise to increase to about −104 dBm. Comparison to the CW tone then gives an input referred IP3 of−20 dBm. Reflecting this to the ADC gives an IP3 of+10 dBm or an intermodulation performance of −74 dBc.

Since this receiver may also be used for reception of a narrowband signal, a quick check of narrowband

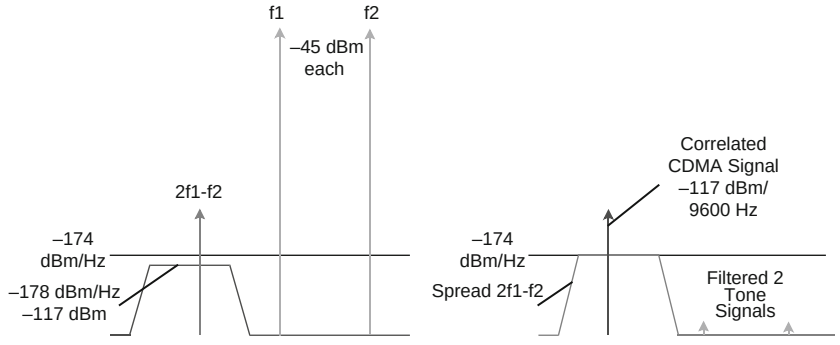


Figure 2.3-16 CDMA2000 two-tone blocking requirements.

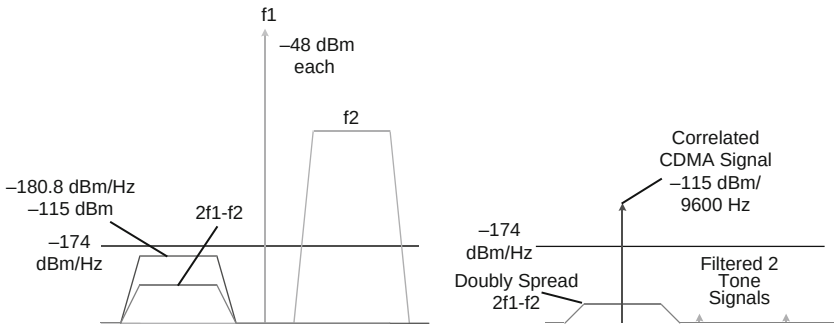


Figure 2.3-17 UMTS intermodulation performance requirements.

performance is a good idea. Total noise referenced to the antenna can be calculated in a 30 kHz band to be −125 dBm. If 5 dB of SNR is required, this is reasonably good performance. In terms of intermodulation rejection, to achieve unrestricted performance at −125 dBm, the intermodulation products from two narrowband terms must be below this level. If the products must be below −130 dBm, and they are generated by a −45 dBm tone, then an input referred IP3 of −2.5 dBm is required, or in terms of single-tone performance with a −30 dBm in band blocker, a −100 dBFS performance. Clearly, from the single-tone requirements, narrowband performance will be limited by the harmonics of the blockers more so than sensitivity.

Clearly, this design maintains good performance for both CDMA2000 and UMTS while retaining the ability to perform reasonably well at narrowband standards. Although this review is not exhaustive, it does indicate a methodology for looking at multimode radio performance. Since this is a wideband receiver, the simultaneous reception in all three modes is possible providing that the digital processing is available (i.e., multi-carrier). Likewise, this receiver is suitable for field configuration between these modes, even if it is not operated in a multimode manner providing many deployment options.

Conclusion

As new and more complex communication standards are developed around the globe, the demand for new transceiver architectures will also grow. However, more and more often the available capital, both financial and human, limits the designs that can be tackled. Fortunately, software radio technology is available for a select and growing group of these architectures that allow a single platform to leverage into many diverse designs. As seen here, this has many distinct advantages and is not limited to interoperability, investment retention, and great flexibility. As with any software project, quite often the potential is only limited by the imagination of the designer. The great advantage is that as in any software project, if there is a design error, it is as simple as backspace, type, and enter to fix the problem.

Fortunately, the last decade has seen significant advances in semiconductor technology that have caused impressive gains [17] not only in performance but also in cost. SDR is one area that has greatly benefited from these varied technologies and will continue to do so as the meaning of SDR is developed, just as has been the case in the history of programming languages.

Although SDR is not the solution to all communication problems, it will offer robust solutions to challenging design issues in the coming years. These issues include phased array technology, location services, interoperability, and complex concepts yet to be defined. However, there are still some challenges preventing full acceptance of this technology. The two main issues are cost and power. Interestingly, these two have a first-order positive relationship; solve one problem and the other will only get better. Without low power, user devices will not be able to take full advantage of SDR technology. Clearly, the power issue comes from the need for high-performance components. High performance means ultra-linear devices. High-linearity devices mean low efficiency through high-standing currents. Therefore, if the issue of how to design high linearity devices with lower power can be solved, and it will, then costs too will also fall, opening the door for many other applications. So the key to continued SDR development and evolution is continued device improvement down the Moore's law curve and continued interest in flexible radio architectures. Despite these challenges, the current state of performance is more than sufficient for engineers and manufacturers to seriously begin to investigate the possibilities of SDR as covered in this text.

References

1. J.H. Reed, *Software Radio: A Modern Approach to Radio Engineering*, Prentice Hall, Upper Saddle River, NJ, 2002.
2. J. Mitola, III, "Software Radio-Cognitive Radio," http://ourworld.compuserve.com/homepages/jmitola/.
3. B. Brannon, D. Efstathiou, and T. Gratzek, "A Look at Software Radios: Are They Fact or Fiction?" *Electronic Design* (December 1998): 117–22.
4. B. Clarke and K. Kreitzer, "Software Radio Concepts," unpublished paper.
5. B. Brannon, "Digital-Radio-Receiver Design Requires Reevaluation of Parameters," *EDN* 43 (November 1998): 163–70.
6. B. Brannon, "New A/D Converter Benefits Digital IFs," *RF Design* 18 (May 1995): 50–65.
7. W.H. Hayward, "Introduction to Radio Frequency Design," The American Radio Relay League, 1994–1996.
8. J.J. Carr, *Secrets of RF Circuit Design* (New York: McGraw-Hill, 2001).
9. B. Brannon, "Fast and Hot: Data Converters for Tomorrow's Software-Defined Radios," *RF Design* 25 (July 2002): 60–66.
10. B. Brannon and C. Cloninger, "Redefining the Role of ADCs in Wireless," *Applied Microwave and Wireless* 13 (March 2001): 94–105.
11. B. Brannon, "DNL and Some of Its Effects on Converter Performance," *Wireless Design and Development* 9 (June 2001): 10.
12. B. Brannon, "Overcoming Converter Nonlinearies with Dither," Analog Devices Applications Note AN-410, www.analog.com.
13. W. Kester, "High-Speed Sampling and High-Speed ADCs," Section 4, High-Speed Design Techniques, www.analog.com.
14. W. Kester, "High-Speed DACs and DDS Systems," Section 6, High-Speed Design Techniques, www.analog.com.
15. About CDMA and CDMA University. Available at: http://www.qualcomm.com.
16. Specifications. Available at: http://www.3gpp2.org.
17. R.H. Walden, "Analog-to-Digital Converter Survey and Analysis," *IEEE Communications Magazine* 17 (April 1999): 539–50.
18. H. Nyquist, "Certain Topics in Telegraph Transmission Theory," *AIEE Transactions* 47 (April 1928): 617–44.
19. AD6645 Datasheet. Available at: http://www.analog.com.

Chapter 2.4a

2.4a

The software defined radio as a platform for cognitive radio

Pablo Robert and Bruce Fette

2.4a.1 Introduction

This chapter explores both the hardware and software domains of software defined radio (SDR). The span of information covered is necessarily broad; therefore, it focuses on some aspects of hardware and software that are especially relevant to SDR design. Beyond their obvious differences, hardware and software analyses have some subtle differences. In general, hardware is analyzed in terms of its capabilities. For example, a particular radio frequency (RF) front-end (RFFE) can transmit up to a certain frequency, a data converter can sample a maximum bandwidth, and a processor can provide a maximum number of million instructions per second (MIPS). Software, in contrast, is generally treated as an enabler. For example, (1) a signal processing library can support types of modulation, (2) an OS can support multithreading, or (3) a particular middleware implementation can support naming structures. Given this general form of viewing hardware and software, this chapter presents hardware choices as an upper bound on performance, and software as a minimum set of supported features and capabilities.

Cognitive radio (CR) assumes that there is an underlying system hardware and software infrastructure that is capable of supporting the flexibility demanded by the cognitive algorithms. In general, it is possible to provide significant flexibility with a series of tunable hardware components that are under the direct control of the cognitive software. In the case of a cognitive system that can support a large number of protocols and air interfaces, it is desirable to have a generic underlying hardware structure.

The addition of a series of generalized computing structures underlying the cognitive engine implies that the cognitive engine must contain hardware-specific knowledge. With this hardware-specific knowledge, the cognitive engine can then navigate the different optimization strategies that it is programmed to traverse. The problem with such knowledge is that a change in the underlying hardware would require a change in the cognitive engine's knowledge base. This problem becomes exacerbated when one considers porting the engine to other radio platforms. For example, there could be a research and development platform that is used to test a variety of cognitive algorithms. As these algorithms mature, it is desirable to begin using these algorithms in deployed systems. Ideally, one would just need to place the cognitive engine in the deployed system's management structure. However, if no abstraction were available to isolate the cognitive engine from the underlying hardware, the cognitive engine would need to be modified to support the new hardware platform. It is clear that an abstraction is desirable to isolate the cognitive engine from the underlying hardware. The abstraction of hardware capabilities for radio software architecture is a primary design issue.

SDR is more than just an abstraction of the underlying hardware from the application. SDR is a methodology for the development of applications, or *waveforms* in SDR parlance, in a consistent and modular fashion such that both software and hardware components can be readily reused from implementation to implementation. SDR also provides the management structure for the description, creation, and tear-down of waveforms. In several respects, SDR offers the same capabilities supported by OSs; SDR is actually a superset of the capabilities provided by an OS. SDR must support a variety of cores, some of which may be deployed simultaneously in the same system. This capability is like a distributed OS designed to run over a heterogenous hardware environment, where heterogenous in this context means not

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2006 Elsevier Inc. All rights reserved

only general purpose processors (GPPs), but also digital signal processors (DSPs), field-programmable gate arrays (FPGAs), and custom computing machines (CCMs). Furthermore, SDR must support the RF and intermediate frequency (IF) hardware that is necessary to interface the computing hardware with radio signals. This support is largely a tuning structure coupled with a standardized interface. Finally, SDR is not a generic information technology (IT) solution in the way that database management is. SDR deals explicitly with the radio domain. This means that context is important. This context is most readily visible in the application programming interface (API), but is also apparent in the strict timing requirements inherent to radio systems, and the development and debugging complexities associated with radio design.

This chapter is organized as follows: Section 2.4a.2 introduces the basic radio hardware architecture and the processing engines that will support the cognitive function. Section 2.4a.3 discusses the software architecture of an SDR. Section 2.4a.4 discusses SDR software design and development. At present, many SDRs utilize a Software Communications Architecture (SCA) as a middleware to establish a common framework for waveforms, and the SCA is covered in some detail in this section. Section 2.4a.5 discusses applications as well as the cognitive functionality and languages that support cognitive software as an application. Section 2.4a.6 discusses the development process for SDR software components. Section 2.4a.7 then discusses cognitive waveform development. Finally, Section 2.4a.8 presents a summary of the chapter.

2.4a.2 Hardware architecture

The underlying hardware structure for a system provides the maximum bounds for performance. The goal of this section is to explore hardware for SDR from a radio standpoint. Figure 2.4a-1(a) shows a basic radio receiver. As one example based on the basic radio receiver architecture, Figure 2.4a-1(b) shows a design choice made possible by digital signal processing techniques, in which the sampling process for digital signal processing can be placed in any of several locations and still provide equivalent performance.

2.4a.2.1 The block diagram

The generic architecture tour presented here traces from the antenna through the radio and up the protocol stack to the application.

RF externals

Many radios may achieve satisfactory performance with an antenna consisting of a passive conductor of resonant length, or an array of conductors that yield a beam pattern. Such antennas range from the simple quarter-wavelength vertical to the multi-element Yagi and its wide bandwidth cousin, the log periodic antenna. Antennas used over a wide frequency range will require an antenna tuner to optimize the voltage standing wave radio (VSWR) and corresponding radiation efficiency. Each time the transceiver changes frequency, the antenna tuner will need to be informed of the new frequency. It will either have a prestored table derived from a calibration process, and then adjust passive components to match the tuning recommendations of the table, or it will sense the VSWR and adapt the tuning elements until a minimum VSWR is attained.

Some modern antennas include a number of passive components spread over the length of the radiating elements that are able to present reasonable VSWR

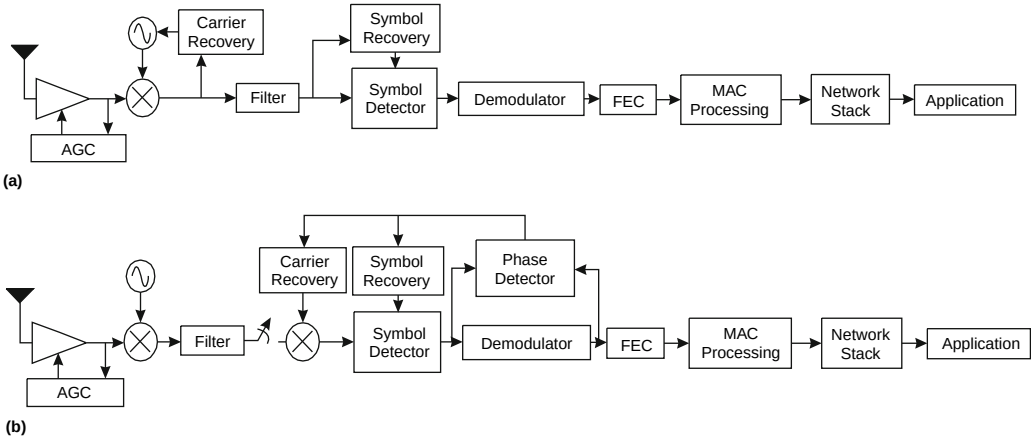


Figure 2.4a-1 (a) Data flow and component structure of a generalized coherent radio receiver. (b) Data flow and component structure of a generalized coherent radio receiver designed for digital systems, with sampling as a discrete step (see *Design Choices* section for an explanation).

performance without active tuning. The best such units today span a frequency range of nearly 10:1. However, for radios that are expected to span 2 MHz–2 GHz, it is reasonable to expect that the radio will need to be able to control a switch to select an appropriate antenna for the frequency band in which the transceiver is currently operating. Where beam antennas are used, it may also be necessary for the radio to be able to manage beam pointing. In some cases, this is accomplished by an antenna rotator, or by a dish gimbals. The logic of how to control the pointing depends greatly on the application. Examples include: (1) exchanging the global positioning system (GPS) position of each transceiver in the network, as well as tracking the three-dimensional (3-D) orientation of the platform on which the antenna is mounted so that the antenna pointing vector to any network member can be calculated; (2) scanning the antenna full circle to find the maximum signal strength; (3) dithering the antenna while tracking peak performance; or (4) using multiple receive feed elements and comparing their relative strength.

Another common antenna is the electronically steered antenna. Control interfaces to these antennas are quite similar in function; however, due to the ability to rapidly steer electronically, many of these antennas update their steering angle as rapidly as once every millisecond. Thus, response time of the control interfaces is critical to the expected function.

The most sophisticated antenna is the multiple input, multiple output (MIMO) antenna. In these antennas, the interface boundary between the radio and the antenna is blurred by the wide bandwidth and large number of complex interfaces between the beam-steering signal processing, the large number of parallel RF front-end receivers, and the final modem signal processing. For these complex antennas, the SDR Forum is developing interface recommendations that will be able to anticipate the wide variety of multi-antenna techniques currently used in modern transceivers.

Another common external component is the RF power amplifier (PA). Typically, the external PA needs to be told to transmit when the transceiver is in the transmit mode and to stop transmitting when the transceiver is in the receive mode. A PA will also need to be able to sense its VSWR, delivered transmit power level, and its temperature, so that the operator can be aware of any abnormal behavior or conditions.

It is also common to have a low-noise amplifier (LNA) in conjunction with an external PA. The LNA will normally have a tunable filter with it. Therefore, it is necessary to be able to provide digital interfaces to the external RF components to provide control of tuning frequency, transmit/receive mode, VSWR and transmit power-level sensing, and receive gain control.

In all of these cases, a general-purpose SDR architecture must anticipate the possibility that it may be called upon to provide one of these control strategies for an externally connected antenna, and so must provide interfaces to control external RF devices. Experience has shown Ethernet to be the preferable standard interface, so that remote control devices for RF external adapters, switches, PAs, LNAs, and tuners can readily be controlled.

RF front-end

The RF front-end consists of the receiver and the transmitter analog functions. The receiver and transmitter generally consist of frequency up converters and down converters, filters, and amplifiers. Sophisticated radios will choose filters and frequency conversions that minimize spurious signals, images, and interference within the frequency range over which the radio must work. The front-end design will also maximize the dynamic range of signals that the receiver can process, through automatic gain control (AGC). For a tactical defense application radio, it is common to be able to continuously tune from 2 MHz to 2 GHz, and to support analog signal bandwidths ranging from 25 kHz to 30 MHz. Commercial applications need a much smaller tuning range. For example, a cell phone subscriber unit might tune only 824 MHz–894 MHz and might need only one signal bandwidth. With a simplified design, the designer can eliminate many filters, frequency conversions, and IF bandwidth filters, with practical assumptions.

The RF analog front-end amplifies and then converts the radio carrier frequency of a signal of interest down to a low IF so that the receive signal can be digitized by an analog-to-digital converter (ADC), and then processed by a DSP to perform the modem function. Similarly, the transmitter consists of the modem producing a digital representation of the signal to be transmitted, and then a digital-to-analog converter (DAC) process produces a baseband or IF representation of the signal. That signal is then frequency shifted to the intended carrier frequency, amplified to a power level appropriate to close the communication link over the intended range, and delivered to the antenna. If the radio must transmit and receive simultaneously, as in full duplex telephony, there will also be some filtering to keep the high-power transmit signal from interfering with the receiver's ability to detect and demodulate the low-power receive signal. This is accomplished by complex filters usually using bulk acoustic wave or saw filters at frequencies below 2 GHz, or yttrium-iron-garnet (YIG) circulators at frequencies above 2 GHz.

The typical software-defined RFFE begins notionally with receiving a signal and filtering the signal to reflect the range of frequency covered by the intended signals. For a spread spectrum wideband code division multiple access (WCDMA) signal, this could be up to 6 MHz of bandwidth. In order to assure that the full 6 MHz is

presented to the modem without distortion, it is not unusual for the ADC to digitize 12 MHz or so of signal bandwidth. In order to capture 12 MHz of analog signal bandwidth set by the IF filters without aliasing artifacts, the ADC will probably sample the signal at rates above 24 million samples per second (Msps). After sampling the signal, the digital circuits will shift the frequency of the RF carrier to be centered as closely as possible to 0 Hz direct current (DC) so that the signal can again be filtered digitally to match the exact signal bandwidth. Usually this filtering will be done with a cascade of several finite impulse response (FIR) filters, designed to introduce no phase distortion over the exact signal bandwidth. If necessary, the signal is despread, and then refiltered to the information bandwidth, typically with an FIR filter.

Analog-to-digital converters

The rate of technology improvement versus time has not been as profound for ADCs as for digital logic. The digital receiver industry is always looking for wider bandwidth and greater dynamic range. Successive approximation ADCs were replaced by flash converters in the early 1990s, and now are generally replaced with sigma-delta ADCs. Today's ADC can provide up to 105 Msps at 14-bit resolution. Special-purpose ADCs have been reported to provide sample rates over 5 giga samples per second (Gsps) at 8-bit resolution.

State-of-the-art research continues to push the boundaries of analog-to-digital (A/D) performance with a wide variety of clever techniques that shift the boundaries between DSP and ADC.

Modem

After down-conversion, filtering, and equalization, the symbols are converted to bits by a symbol detector/demodulator combination, which may include a matched filter or some other detection mechanism as well as a structure for mapping symbols to bits. A symbol is selected that most closely matches the received signal. At this stage, timing recovery is also necessary, but for symbols rather than the carrier. Then the output from the demodulator is in bits.

The bits that are represented by that symbol are then passed to the forward error correcting function to correct occasional bit errors. Finally, the received and error-corrected bits are parsed into the various fields of message, header, address, traffic, etc. The message fields are then examined by the protocol layers eventually delivering messages to an application (e.g., Web browser or voice coder (VoCoder)), thus delivering the function expected by the user.

SDRs must provide a wide variety of computational resources to be able to gracefully anticipate the wide variety of waveforms; they may ultimately be expected to demodulate. Today we would summarize that a typical SDR should be able to provide at least 266 MIPS of GPP, 32 Mbytes of random access memory (RAM), 100 MIPS of DSP, and 500 K equivalent gates of FPGA-configurable logic. More performance and resources are required for sophisticated waveforms or complex networking applications. Typically, the GPP will be called upon to perform the protocol stack and networking functions, the DSP will perform the physical layer modulation and demodulation, the FPGA will provide timing and control as well as any special-purpose hardware accelerators that are particularly unique to the waveform. It appears that SDR architectures will continue to evolve as component manufacturers bring forward new components that shift the boundaries of lowest cost, highest performance, and least power dissipation.

Forward error correction

In some instances, the demodulated bits are passed on to a forward error correction (FEC) stage for a reduction in the number of bit errors received. One of the interesting aspects of FEC is that it can be integrated into the demodulation process, such as in trellis-coded modulation; or it can be closely linked to demodulation, as in soft decoding for convolutional codes; or it can be an integral part of the next stage, medium access control (MAC) processing.

Medium access control

MAC processing generally includes framing information, with its associated frame synchronization structures, MAC addressing, error detection, link management structures, and payload encapsulation with possible fragmentation/defragmentation structures. From this stage, the output is bits, which are input to the network-processing layer. The network layer is designed for end-to-end connectivity support. The output of the network layer is passed to the application layer, which performs some sort of user functions and interface (speaker/microphone, graphical user interface, keypad, or some other sort of human-computer interface).

User application

The user's application may range from voice telephony, to data networking, to text messaging, to graphic display, to live video. Each application has its own unique set of requirements, which, in turn, translate into different implications on the performance requirements of the SDR.

For voice telephony today, the dominant mode is to code the voice to a moderate data rate. Data rates from 4800 bps to 13,000 bps are popular in that they provide

excellent voice quality and low distortion to the untrained listener. The digital modem, in turn, is generally more robust to degraded link conditions than analog voice would be under identical link conditions.

Another criterion for voice communications is low latency. Much real experience with voice communications makes it clear that if the one-way delay for each link exceeds 50 milliseconds,[1] then users have difficulty in that they expect a response from the far speaker and, hearing none, they begin to talk just as the response arrives, creating frequent speech collisions. In radio networks involving ad hoc networking, due to the delay introduced by each node as it receives and retransmits the voice signaling, it can be quite difficult to achieve uniformly low delay. Since the ad hoc network introduces jitter in packet delivery, the receiver must add a jitter buffer to accommodate a practical upper bound in latency of late packets. All of this conspires to add considerable voice latency. In response, voice networks have established packet protocols that allocate traffic time slots to the voice channels, in order to guarantee stable and minimal latency. In much the same way, video has both low error rate and fixed latency channel requirements, and thus networking protocols have been established to manage the quality requirements of video over wireless networks. Many wireless video applications are designed to accept the bit errors but maintain the fixed latency.

In contrast, for data applications, the requirement is that the data must arrive with very few or absolutely no bit errors; however, latency is tolerated in the application.

Voice coding applications are typically implemented on a DSP. The common voice coding algorithms require between 20 and 60 MIPS and about 32 Kbytes of RAM. Voice coding can also be successfully implemented on GPPs, and will typically require more than six times the instructions per second (100–600 MIPS) in order to perform both the multiply—accumulate signal processing arithmetic and the address operand fetch calculations.

Transmitting video is nearly 100 times more demanding than voice, and is rarely implemented in GPP or DSP. Rather, video encoding is usually implemented on special purpose processors due to the extensive cross-correlation required to calculate the motion vectors of the video image objects. Motion vectors substantially reduce the number of bits required to faithfully encode the images. In turn, a flexible architecture for implementing these special purpose engines is the use of FPGAs to implement the cross-correlation motion-detection engines.

Web browsing places a different type of restriction on an SDR. The typical browser needs to be able to store the images associated with each Web page in order to make the browsing process more efficient, by eliminating the redundant transmission of pages recently seen. This implies some large data cache function, normally implemented by a local hard drive disk. Recently, such memories are implemented by high-speed flash memory as a substitute for rotating electromechanical components.

Design choices

Several aspects of the receiver shown in Figure 2.4a-1(a) are of interest. One of the salient features is that the effect of all processing between the LNA and the FEC stage can be largely modeled linearly. This means that the signal processing chain does not have to be implemented in the way shown in Figure 2.4a-1(a). The carrier recover loop does not have to be implemented at the mixer stage. It can just as easily be implemented immediately before demodulation. Another point to note is that no sampling is shown in Figure 2.4a-1(a). It is mathematically possible to place the sampling process anywhere between the LNA and the FEC, giving the designer significant flexibility. An example of such design choice selection is shown in Figure 2.4a-1(b), where the sampling process is shown as a discrete step.

The differences seen between Figures 2.4a.1(a) and 2.4a.1(b) are not at a functional level; they are implementation decisions that are likely to lead to a dramatically different hardware structure for equivalent systems. The following discussion on hardware concentrates on processing hardware selections because a discussion of RF and IF design considerations is beyond the scope of this book.

Several key concepts must be taken into consideration for the front-end of the system, and it is worthwhile to briefly mention them here. From a design standpoint, signals other than the signal of interest can inject more noise in the system than was originally planned, and the effective noise floor may be significantly larger than the noise floor due to the front-end amplifier.

One example of unpredicted noise injection is the ADC conversion process, which can inject noise into the signal through a variety of sources. The ADC quantization process injects noise into a signal. This effect becomes especially noticeable when a strong signal that is not the signal of interest is present in an adjacent channel. Even though it will be removed by digital filtering, the stronger signal sets the dynamic range of the receiver by affecting the AGC, in an effort to keep the ADC from being driven into saturation.

[1] Some systems specify a recommended maximum latency limit, such as 150 milliseconds for ITU-T G114.

Thus, the effective signal-to-interference and noise ratio (SINR) of the received signal is lower than might otherwise be expected from the receiver front-end. To overcome this problem and similar ones, software solutions will not suffice, and flexible front-ends that are able to reject signals before sampling become necessary. Tunable RF components, such as tunable filters and amplifiers, are becoming available, and the SDR design that does not take full advantage of these flexible front-ends will handicap the final system performance.

2.4a.2.2 Baseband processor engines

The dividing line between baseband processing and other types of processing, such as network stack processing, is arbitrary, but it can be constrained to be between the sampling process and the application. The application can be included in this portion of processing, such as a VoCoder for a voice system or image processing in a video system. In such instances, the level of signal processing is such that it may be suitable for specialized signal processing hardware, especially in demanding applications such as video processing.

Four basic classes of programmable processors are available today: GPPs, DSPs, FPGAs, and CCMs.

General purpose processors

GPPs are the target processors that probably come to mind first to anyone writing a computer program. GPPs are the processors that power desktop computers and are at the center of the computer revolution that began in the 1970s. The landscape of microprocessor design is dotted with a large number of devices from a variety of manufacturers. These different processors, while unique in their own right, do share some similarities, namely, a generic instruction set, an instruction sequencer, and a memory management unit (MMU).

There are two general types of instruction sets: (1) machines with fairly broad instruction sets, known as complex instruction set computers (CISCs); and (2) machines with a narrow instruction set, known as reduced instruction set computers (RISCs). Generally, the CISC instructions give the assembly programmer powerful instructions that address efficient implementation of certain common software functions. RISC instruction sets, while narrower, are designed to produce efficient code from compilers. The differences between the CISC and RISC are arbitrary, and both styles of processors are converging into a single type of instruction set. Regardless of whether the machine is CISC or RISC, they both share a generic nature to their instructions. These include instructions that perform multiplication, addition, or storage, but these instruction sets are not tailored to a particular type of application. In the context of CR, the application in which we are most interested is signal processing.

The other key aspect of the GPP is the use of an MMU. Because GPPs are designed for generic applications, they are usually coupled with an OS. This OS creates a level of abstraction over the hardware, allowing the development of applications with little or no knowledge of the underlying hardware. Management of memory is a tedious and error-prone process, and in a system running multiple applications, memory management includes paging memory, distributed programming and data storage throughout different blocks of memory. An MMU allows the developer to "see" a contiguous set of memory, even though the underlying memory structure may be fragmented or too difficult to control in some other fashion (especially in a multitasking system that has been running continuously for an extended period of time). Given the generic nature of the applications that run on a GPP, an MMU is critical because it allows the easy blending of different applications with no special care needed on the developer's part.

Digital signal processors

DSPs are specialized processors that have become a staple of modern signal processing systems. In large part, DSPs are similar to GPPs. They can be programmed with a high-level language such as C or C++ and they can run an OS. The key difference between DSPs and GPPs comes in the instruction set and memory management. The instruction set of a DSP is customized to particular applications.

For example, a common signal processing function is a filter, an example of which is the Direct Form II infinite impulse response (IIR) filter. Such a filter is seen in Figure 2.4a-2.

As seen in Figure 2.4a-2, the signal path is composed of delays on the incoming signal of one sample, z^{-1}, and the multiplication of each delayed sample by a coefficient of the polynomial describing either the poles (a) or zeros (b) of the filter. If each delayed sample is considered to be a different memory location, then to quickly implement this filter, it is desirable to perform a sample shift in the circular buffer, perform a multiply and an add that multiplies each delayed sample times the corresponding polynomial coefficients, and store that result either in the output register, in this case $y[n]$, or into the register that is added to the input, $x[n]$.

The algorithm in Figure 2.4a-2 has several characteristics. Assuming that the filter is of length N (N-order polynomials), then the total computation cost for this algorithm can be computed. To optimize radio modem performance, filters are frequently designed to

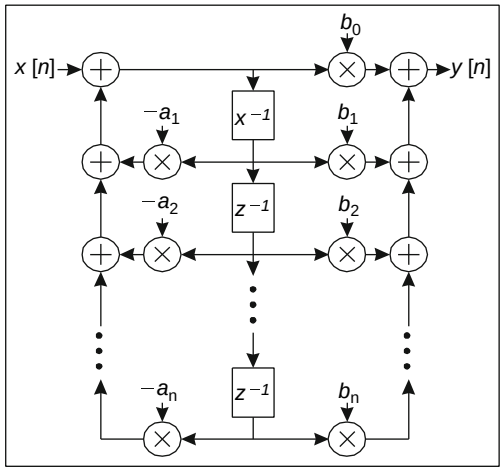


Figure 2.4a-2 Structure and data flow for Direct Form II IIR filter as a basis for an estimate on computational load.

be FIR filters with only the b (all zeros) polynomial. In order to implement extremely efficient DSP architectures, most DSP chips support performing many operations in parallel to make an FIR filter nearly a single instruction that is executed in a 1 clock cycle instruction loop. First, there is a loop control mechanism. This loop control has a counter that has to be initialized and then incremented in each operation, providing a set of $N+1$ operations. Within the loop, there is also an evaluation of the loop control state; this evaluation is performed N times. Within the loop, a series of memory fetch operations have to be performed. In this case, there are $N+1$ accesses to get the coefficient and store the result. There are additional $N+1$ circular accesses of the sample data plus the fetch operation of the new data point, yielding a total of $2N+2$ memory accesses. Finally, there are the arithmetic operations: the coefficient multiplied by the signal sample, and the accumulation operation (including the initial zero operation) resulting in $2N+1$ operations. Therefore, the DSP typically performs on the order of $6N$ operations per instruction clock cycle.

Assuming that a GPP performs a memory fetch, memory store, index update, comparison, multiplication, or addition in each computer clock cycle, then the GPP would require $6N+3$ clock cycles per signal sample. Using these assumptions, then an FIR filter with 32 filter taps and a signal sampled at l00 ksps would yield $(6 \times 32 + 3) \cdot 100 \times 10^3 = 19.5$ MIPS. Therefore, just for the filter mentioned, almost 20 MIPS are required for the GPP to successfully filter this narrowband signal. In reality, a GPP is likely to expend more than one cycle for each of these operations. For example, an Intel Pentium 4 floating point multiply occupies 6 clock cycles, so the given performance is a lower bound on the GPP MIPS load.

A DSP, in contrast, has the ability to reduce some of the cycles necessary to perform the given operations. For example, some DSPs have single-cycle MAC (multiple and accumulate). The Motorola 56001 is an example DSP that performs single instruction multiply and accumulate with zero overhead looping. These reductions result in a computational total of $3N+3$. Given these reductions, the computation load for the DSP is now $(3 \times 32 + 3) \cdot 100 \times 10^3 = 9.9$ MIPS.

Given its customized instruction set, a DSP can implement specialized signal processing functionality with a significantly fewer clock cycles than the more generic GPP processors. However, GPPs are attempting to erode this difference using multiple parallel execution arithmetic logic units so that they can perform effective address calculations in parallel with arithmetic operations. These are called superscalar architectural extensions. They are also attempting to raise the performance of GPP multipliers through use of many pipeline stages, so that the multiplication steps can be clocked at higher speed. This technique is called a highly pipelined architecture.

Field-programmable gate arrays

FPGAs are programmable devices that are different in nature from GPPs and DSPs. An FPGA comprises some discrete set of units, sometimes referred to as logical elements (LE), logic modules (LM), slices, or some other reference to a self-contained Boolean logical operation. Each of these logical devices has at least one logic unit; this logic unit could be one or more multipliers and one or more accumulators, or a combination of such units in some FPGA chip selections. Logical devices are also likely to contain some memory, usually a few bits. The developer then has some freedom to configure each of these logical devices, where the level of reconfigurability is arbitrarily determined by the FPGA manufacturer. The logical devices are set in a logic fabric, a reconfigurable connection matrix that allows the developer to describe connections between different logical devices. The logic fabric usually also has access to some additional memory that logical devices can share. Timing of the different parts of the FPGA can be controlled by establishing clock domains. This allows the implementation of multirate systems on a single FPGA.

To program an FPGA, the developer describes the connections between logical devices as well as the configuration of each of these logical devices. The final design that the developer generates is closer to a circuit than a program in the traditional sense, even though the FPGA is ostensibly a firmware programmable device. Development for the FPGA is done by using languages such as very highspeed integrated circuit (VHSIC)

Hardware Design Language (VHDL), which can also be used to describe application-specific integrated circuits (ASICs), essentially non-programmable chips. Variants of C exist, such as System-C, that allow the developer to use C-like constructs to develop FPGA code, but the resulting program still describes a logic circuit running on the FPGA.

The most appealing aspect of FPGAs is their computational power. For example, the Virtex II Pro FPGA by Xilinx has a total of 6460 slices, where each slice is composed of two look-up tables, two flip-flops, some math logic, and some memory. To be able to implement 802.11a, a communications standard that is beyond the abilities of any traditional DSP in 2005, would require approximately 3000 slices, or less than 50 percent of the FPGA's capabilities, showing the importance of a high degree of parallelism in the use of many multiply accumulators to implement many of the complex-waveform signal processes in parallel.

From a performance standpoint, the most significant drawback of an FPGA is that it consumes a significant amount of power, making it impractical for battery powered handheld subscriber solutions. For example, the Virtex II Pro FPGA mentioned above is rated at 2648 mW of power expenditure, whereas a low-power DSP such as the TMS320C55x is rated at 65–160 mW, depending on clock speed and version, and its high-performance cousin, the TMS320C64x, is rated at 250–1650 mW, depending on clock speed and version.

2.4a.2.3 Baseband processing deployment

One of the problems that the CR developer will encounter when designing a system is attempting to determine what hardware should be included in the design and what baseband processing algorithm to deploy into which processors. The initial selection of hardware will impose limits on maximum waveform capabilities, while the processing deployment algorithm will present significant run-time challenges if left entirely as an optimization process for the cognitive algorithm. If the processing deployment algorithm is not designed correctly, it may lead to a sub-optimal solution that, while capable of supporting the user's required quality of service (QoS), may not be power efficient, quickly draining the system batteries or creating a heat dissipation problem.

The key problem in establishing a deployment methodology is one of scope. Once a set of devices and algorithm performance for each of these devices has been established, there is a finite set of possibilities that can be optimized. The issue with such optimization is not the optimization algorithm itself; several algorithms exist today for optimizing for specific values, such as the minimum mean square error (MMSE), maximum likelihood estimation (MLE), genetic algorithms, neural nets, or any of a large set of algorithms. Instead, the issue is in determining the sample set over which to perform the optimization.

There is no openly available methodology for establishing the set of possible combinations over which optimization can occur. However, at least one approach has been suggested by Neel et al. [1] that may lead to significant improvements in deployment optimization. The proposed methodology is partitioned into platform-specific and waveform-specific analyses. The platform-specific analysis is further partitioned into two types, DSP/GPP and FPGA. The platform-specific analysis is as follows:

1. Create an operations audit of the target algorithms (number and type of operations).
2. For DSP:
   - **(a)** Create a set of target devices.
   - **(b)** Establish cycle-saving capabilities of each target device.
3. For FPGA:
   - **(a)** Create a set of devices.
   - **(b)** Establish mapping between different FPGA families. This mapping can be done on the basis of logical devices, available multiplies per element, or another appropriate metric.
   - **(c)** Find logical device count for each target algorithm. FPGA manufacturers usually maintain thorough libraries with benchmarks.
   - **(d)** Use mapping between devices to find approximate target load on devices when a benchmark is not available.

Once the platform-specific analysis is complete, the developer now has the tools necessary to map specific algorithms onto a set of devices. Note that at this stage, there is no information as to the suitability of each of these algorithms to different platforms, since a base clock rate (or data rate) is required to glean that type of information.

Given the platform-specific information assembled above, it is now possible to create performance estimates for the different waveforms that the platform is intended to support:

1. Create a block-based breakdown of the target waveform (using target algorithms from step 1 as building blocks).
2. Breakdown target waveform into clock domains.

3. Estimate time necessary to complete each algorithm.
   (a) In the case of packet-based systems, this value is fairly straightforward.[2]
   (b) In the case of stream-based systems, this value is the allowable latency.
4. Compute number of operations per second (OPS) needed for each algorithm.
5. Create a set of devices from the platform-specific phase that meet area and cost parameters (or whatever other parameters are appropriate for a first cut). This set of devices can be very large. At this stage, the goal is to create a set of devices or combination of devices that meet some broad criteria.
6. Cycle through the device set.
   (a) Attempt to map algorithms onto given devices in set.
      (i) For DSP:
         (1) Make sure that OPS calculated in step 4 of the waveform-specific analysis are reduced by cycle-saving capabilities outlined in step 2b of the platform-specific analysis.
         (2) The result of the algorithm map is an MIPS count for each device.
      (ii) For FPGA:
         (1) Mapping of the algorithm is a question of the number of occupied logical devices.
         (2) Make sure that clock domains needed for algorithms can be supported by the FPGA.
   (b) If a solution set of MIPS and/or LE exists for the combination of devices in the set, then save the resulting solution set for this device set; if a solution does not exist, discard the device set.
7. Apply appropriate optimization algorithm over resulting solution set/device set from step 6. Additional optimization algorithms include power budgets and performance metrics.

The process described yields a coarse solution set for the hardware necessary to support a particular set of baseband processing solutions. From this coarse set, traditional tools can be used to establish a more accurate match of resources to functionality. Furthermore, the traditional tools can be used to find a solution set that is based on optimization criteria that are more specific to the given needs.

2.4a.2.4 Multicore systems and system-on-chip

Even though several computing technologies have some promise over the horizon, such as quantum computing, it is an undeniable fact that silicon-based computing such as multicore systems and system-on-chip (SoC) will continue to be the bedrock of computing technology. Unfortunately, as technology reaches transistors under 100 nm, the key problems become the inability to continue the incremental pace of clock acceleration as well as significant problems in power dissipation. Even though the number of gates per unit area has roughly doubled every 18 months since the 1970s, the amount of power consumed per unit area has remained unchanged. Furthermore, the clocks driving processors have reached a plateau, so increases in clock speed have slowed significantly.

In order to overcome the technology problems in fabrication, a design shift has begun in the semiconductor industry. Processors are moving away from single-core solutions to multicore solutions, in which a chip is composed of more than one processing core. Several advantages are evident from such solutions. First, even though the chip area is increasing, it is now populated by multiple processors that can run at lower clock speeds. In order to understand the ramifications of such change, it is first important to recall the power consumption of an active circuit as

$$P = \alpha \cdot C \cdot f \cdot V^2 \tag{2.4a.1}$$

As shown in Eq. (2.4a.1), the power dissipated, P, by an active circuit is the product of the switching activity, α, the capacitance of the circuit, C, the clock speed, f, and the operating voltage, V, squared. It is then clear from the above equation that the reduced clock speed results in a proportional reduction in power consumption. Furthermore, since a lower operating frequency means that a lower voltage is needed to operate the device, the reduction in the operating voltage produces a reduction in power consumption that follows a quadratic curve.

One of the principal bottlenecks in processor design is the input/output interface from data inside the chip to circuits outside the chip. This interface tends to be significantly slower than the data buses inside the chip. By reducing this interface capacitance and voltage swing for intercore communications, system efficiency grows. Furthermore, communication through shared memory is

[2] If the received signal is blocked into a block of many signal samples, and then the receiver operates on that block of signal samples through all of the receive signal processes, the process can be imagined to be a signal packet passing through a sequence of transforms. In contrast, if each new signal sample is applied to the entire receive process that is referred to here as a stream-based process.

now possible within the chip. This capability can greatly increase the efficiency of the design.

2.4a.3 Software architecture

Software is subject to differences in structure that are similar to those differences seen in the hardware domain. Software designed to support baseband signal processing generally does not follow the same philosophy or architecture that is used for developing application-level software. Underlying these differences is the need to accomplish a variety of quite different goals. This section outlines some of the key development concepts and tools that are used in modern SDR design.

2.4a.3.1 Design philosophies and patterns

Software design has been largely formalized into a variety of design philosophies, such as object-oriented programming (OOP), component-based programming (CBP), or aspect-oriented programming (AOP). Beyond these differences is the specific way in which the different pieces of the applications are assembled, which is generally referred to as a design pattern. This section describes design philosophies first, providing a rationale for the development of different approaches. From these philosophies, the one commonly used in SDR will be expanded into different design patterns, showing a subset of approaches that are possible for SDR design.

Design philosophies

Four basic design philosophies are used for programming today: linear programming (LP), OOP, CBP, and AOP.

Linear programming

LP is a methodology in which the developer follows a linear thought process for the development of the code. The process follows a logical flow, so this type of programming is dominated by conditional flow control (such as "if–then" constructs) and loops. Compartmentalized functionality is maintained in functions, where execution of a function involves swapping out the stack, essentially changing the context of operation, performing the function's work, and returning results to the calling function, which requires an additional stack swap. An analogy of LP is creating a big box for all items on your desktop, such as the phone, keyboard, mouse, screen, headphone, can of soda, and picture of your attractive spouse, with no separation between these items. Accessing any one item's functionality, such as drinking a sip of soda, requires a process to identify the soda can, isolate the soda can from the other interfering items, remove it from the box, sip it, and then place it back into the box and put the other items back where they were. C is the most popular LP language today, with assembly development reserved for a few brave souls who require truly high speed without the overhead incurred by a compiler.

Object-oriented programming

OOP is a striking shift from LP. Whereas LP has data structures—essentially variables that contain an arbitrary composition of native types such as float or integer—OOP extends the data structure concept to describe a whole object. An object is a collection of member variables (such as in a data structure) and functions that can operate on those member variables. From a terminology standpoint, a class is an object's type, and an object is a specific instance of a particular class. There are several rules governing the semantics of classes, but they generally allow the developer to create arbitrary levels of openness (or visibility), different scopes, different contexts, and different implementations for function calls that have the same name. OOP has several complex dimensions; additional information can be found elsewhere (e.g., Budd [2] and Weisfeld [3]).

The differences inherent in OOP have dramatic implications for the development of software. Extending the analogy from the previous example, it is now possible to break up every item on your desktop into a separate object. Each object has some properties, such as the temperature of your soda, and each object also has some functions that you can access to perform a task on that object, such as drinking some of your soda. There are several languages today that are OOP languages. The two most popular ones are Java and C++, although several other languages today are also OOP languages.

Component-based programming

CBP is a subtle extension of the OOP concept. In CBP, the concept of an object is constrained; instead of allowing any arbitrary structure for the object, under CBP the basic unit is now a component. This component comprises one or more classes, and is completely defined by its interfaces and its functionality. Again extending the previous example, the contents on the desktop can now be organized into components. A component could be a computer, where the computer component is defined as the collection of the keyboard, mouse, display, and the actual computer case. This particular computer component has two input interfaces (the keyboard and the mouse) and one output interface (the display). In future generations of this component, there could be additional interfaces, such as a set of headphones as an output interface, but the component's legacy interfaces are not affected by this new capability. Using CBP, the nature of the computer is irrelevant to the user as long as the interfaces and functionality remain the same. It is now possible to change individual objects within the component, such as the keyboard, or the whole

component altogether, but the user is still able to use the computer component the same as always. The primary goal of CBP is to create stand-alone components that can be easily interchanged between implementations. Note that CBP is a coding style, and there are no mainstream languages that are designed explicitly for CBP.

Even though CBP relies on a well-defined set of interfaces and functionality, these aspects are insufficient to guarantee that the code is reusable or portable from platform to platform. The problem arises not from the concept, but from the implementation of the code. To see the problem, it is important now to consider writing the code describing the different aspects of the desktop components that we described before, in this case a computer. Conceptually, we have a component that contains an instance of a display, keyboard, mouse, headphone, and computer. If one were to write software emulating each of these items, not only would the interfaces and actual functional specifications need to be written, but also a wide variety of housekeeping functions, including, for example, notification of failure. If any one piece of the component fails, it needs to inform the other pieces that it failed, and the other pieces need to take appropriate action to prevent further malfunctions. Such a notification is an inherent part of the whole component, and implementing changes in the messaging structure for this notification on any one piece requires the update of all other pieces that are informed of changes in state. These types of somewhat hidden relationships create a significant problem for code reuse and portability because relationships that are sometimes complex need to be verified every time that code is changed. AOP was designed to attempt to resolve this problem.

Aspect-oriented programming

AOP allows for the creation of relationships between different classes. These relationships are arbitrary, but can be used to encapsulate the housekeeping code that is needed to create compatibility between two classes. In the messaging example, this class can include all the messaging information needed for updating the state of the system. Given this encapsulation, a class such as the headphone in the ongoing example can be used not only in the computer example, but also in other systems, such as a personal music player, an interface to an airplane sound system, or any other appropriate type of system. The relationship class encompasses an aspect of the class; thus, context can be provided through the use of aspects. Unlike CBP, AOP requires the creation of new language constructs that can associate an aspect to a particular class; to this end, there are several languages (AspectJ, AspectC++, and Aspect#, among others).

Design philosophy and SDR

The dominant philosophy in SDR design is CBP because it closely mimics the structure of a radio system, namely the use of separate components for the different functional blocks of a radio system, such as link control or the network stack. SDR is a relatively new discipline with few open implementation examples, so as the code base increases and issues in radio design with code portability and code reuse become more apparent, other design philosophies may be found to make SDR software development more efficient.

Design patterns

Design patterns are programming methodologies that a developer uses within the bounds of the language the developer happens to be using. In general, patterns provide two principal benefits: they help in code reuse and they create a common terminology. This common terminology is of importance when working on teams because it simplifies communications between team members. As will be shown in the next section, some architectures, such as the SCA, use patterns. For example, the SCA uses the factory pattern for the creation of applications. In the context of this discussion, patterns for the development of waveforms and the deployment of cognitive engines will be shown. The reader is encouraged to explore formal patterns using available sources (e.g., Gamma et al. [4], Shalloway and Trott [5], or Kerievsky [6]).

2.4a.4 SDR development and design

From the discussion above, it is clear that a software structure following a collection of patterns is needed for efficient large-scale development. In the case of SDR, the underlying philosophy coupled with a collection of patterns is called an architecture or operating environment. There are two open SDR architectures, GNURadio and SCA.

2.4a.4.1 GNURadio

GNURadio [7] is a Python-based architecture (see section *Python)* that is designed to run on general-purpose computers running the Linux OS. GNURadio is a collection of signal processing components and supports primarily one RF interface, the universal software radio peripheral (USRP), a four channel up- and down-converter board coupled with ADC and DAC capabilities. This board also allows the use of daughter RF boards. GNURadio in general is a good starting point for entry-level SDR and

should prove successful in the market, especially in the amateur radio and hobbyist market. GNURadio does suffer from some limitations—namely, (1) that it is reliant on GPP for baseband processing, thus limiting its signal processing capabilities on any one processor, and (2) it lacks distributed computing support, limiting solutions to single-processor systems, and hence limiting its ability to support high-bandwidth protocols.

2.4a.4.2 Software communications architecture

The other open architecture is the SCA, sponsored by the Joint Program Office (JPO) of the US Department of Defense (DoD) under the Joint Tactical Radio System (JTRS) program. The SCA is a relatively complex architecture that is designed to provide support for secure signal processing applications running on heterogenous, distributed hardware. Furthermore, several solutions are available today that provide support for systems using this architecture, some of which are available openly, such as Virginia Tech's OSSIE [8] or Communications Research Center's SCARI [9], providing the developer with a broad and growing support base.

The SCA is a component management architecture; it provides the infrastructure to create, install, manage, and de-install waveforms as well as the ability to control and manage hardware and interact with external services through a set of consistent interfaces and structures. There are some clear limits to what the SCA provides. For example, the SCA does not provide such real-time support as maximum latency guarantees or process and thread management. Furthermore, the SCA also does not specify how particular components must be implemented, what hardware should support what type of functionality, or any other deployment strategy that the user or developer may follow. The SCA provides a basic set of rules for the management of software on a system, leaving many of the design decisions up to the developer. Such an approach provides a greater likelihood that the developer will be able to address the system's inherent needs.

The SCA is based on some underlying technology to be able to fulfill two basic goals, namely, code portability and reuse. In order to maintain a consistent interface, the SCA uses Common Object Request Broker Architecture (CORBA) as part of its middleware. CORBA is software that allows a developer to perform remote procedure calls (RPCs) on objects as if they resided in the local memory space even if they reside in some remote computer. The specifics of CORBA are beyond the scope of this book, but a comprehensive guide on the use of CORBA is available [10].

In recent years, implementations of CORBA have appeared on DSP and FPGA, but traditionally CORBA has been written for GPP. Furthermore, system calls are performed through an OS, requiring an OS on the implementation. In the case of the SCA, the OS of choice is a portable operating system interface (POSIX) PSE-52 compliant OS, but CORBA need not limit itself to such an OS. Its GPP-centric focus leads to a flow for the SCA, as seen in Figure 2.4a-3.

As seen in Figure 2.4a-3, at the center of the SCA implementation is an OS, implying but not requiring the use of a GPP. Different pieces of the SCA are linked to this structure through CORBA and the interface definition language (IDL). IDL is the language used by CORBA to describe component interfaces, and is an integral part of CORBA. The different pieces of the system are attached together by using IDL. An aspect of the SCA that is not obvious from Figure 2.4a-3 is that there can be more than one processor at the core, since CORBA provides location independence to the implementation. Beyond this architecture constraint are the actual pieces that make up the functioning SCA system, namely SCA and legacy software, non-CORBA processing hardware, security, management software, and an integrated file system.

The SCA is partitioned into four parts: the framework, the profiles, the API, and the waveforms. The framework is further partitioned into three parts, base components, framework control, and services. Figure 2.4a-4 is a diagram of the different classes and their corresponding parts.

The SCA follows a component-based design, so the whole infrastructure revolves around the goal of creating, installing, managing, and de-installing the components making up a particular waveform.

Base components

The base class of any SCA-compatible component is the Resource class. The Resource class is used as the parent class for any one component. Since components are described in terms of interfaces and functionality, not the individual makeup of the component, it follows that any one component comprises one or more classes that inherit from the Resource. For example, the developer may create a general Filter category for components. In this example, the filter may be implemented as an FIR or IIR filter. These filters can then be further partitioned into specific Filter implementations up to the developer's discretion, but for the purposes of this example, assume that the developer chooses to partition the filters only into finite or infinite response times (rather than some other category such as structure (e.g., Butterworth, Chebyshev) or spectral response (e.g., low-pass, high-pass, notch). If the developer chooses to partition the

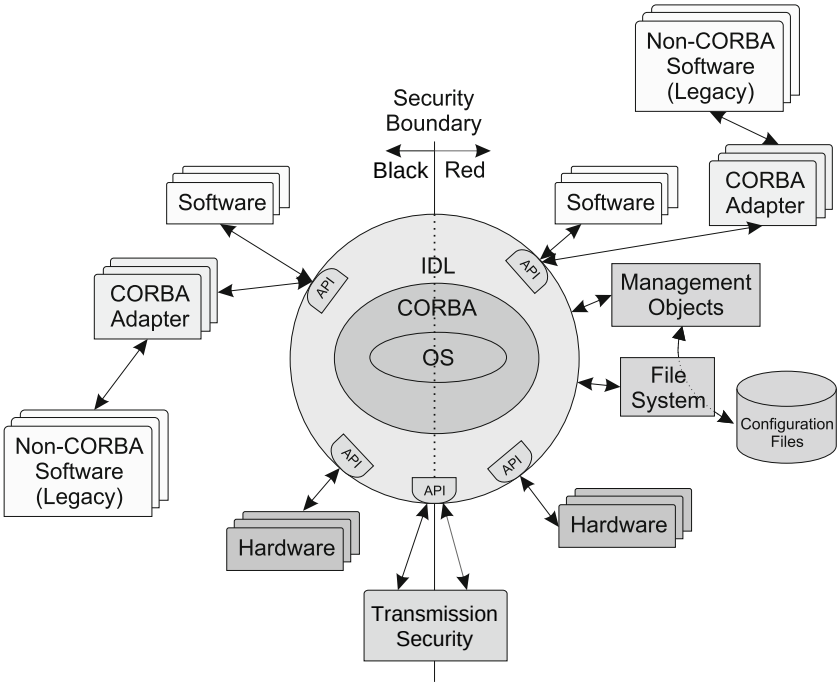


Figure 2.4a-3 CORBA-centric structure for the SCA. The DoD implementation of the SCA requires that the RF modem (black) side of an SDR be isolated from the side where the plain text voice or data is being processed (red), and that this isolation be provided by the cryptographic device.

Filter implementation into FIR and IIR, then a possible description of the Filter family of classes can be that shown in Figure 2.4a-5.

The Resource base class has only two member methods (start() and stop()) and one member attribute (identifier). A component must do more than just start and stop, it must also be possible to set the component's operational parameters, initialize the component into the framework and release the component from the framework, do some sort of diagnostic, and connect this

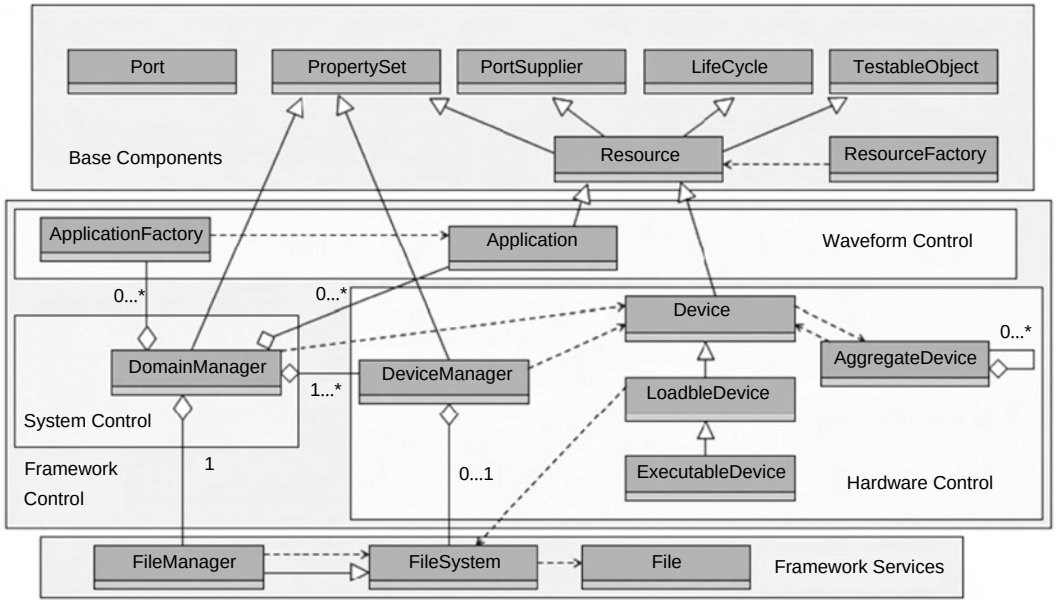


Figure 2.4a-4 Classes making up the SCA core framework.

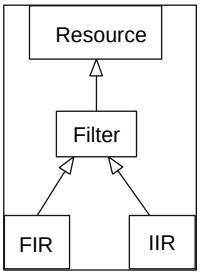


Figure 2.4a-5 Sample filter component class hierarchy.

component to other components. In order to provide this functionality, Resource inherits from the following classes: PropertySet, LifeCycle, TestableObject, and PortSupplier, as seen in Figure 2.4a-6.

PropertySet

The PropertySet class provides a set of interfaces that allow other classes to both configure and query the values associated with this Resource. As part of the creation process of the component, the framework reads a configuration file and sets the different values associated with the component through the PropertySet interface. Later on, during the run-time behavior of the component, the Resource's properties can be queried through the query( ) interface and reconfigured with the configure( ) interface.

LifeCycle

The LifeCycle parent class provides the Resource with the ability to both initialize and release the component from the framework through the initialize( ) and releaseObject( ) interfaces. Initialization in this context is not the same as configuring the different values associated with the component. In this context, initialization sets the component to a known state. For example, in the case of a filter, the initialization call may allocate the memory necessary to perform the underlying convolution and it may populate the filter polynomial from the component's set of properties. The releaseObject performs the complimentary function to initialize. In the case of a filter, it would deallocate the memory needed by the component. One aspect of releaseObject that is common to all components is that it unbinds the component from CORBA. In short, releaseObject performs all the work necessary to prepare the object for destruction.

TestableObject

Component test and verification can take various forms, so it is beyond the scope of the SCA to outline all tests that are possible with a component. However, the SCA does provide a simple test interface through the use of the TestableObject parent class. TestableObject contains a single interface, runTest( ). runTest( ) takes as an input parameter a number signifying the test to be run and a property structure to provide some testing parameters, allowing the inclusion of multiple tests in the component. Even though the interface provides the ability to support multiple tests, it is fundamentally a black box test structure.

PortSupplier

The final capability that parent classes provide to a Resource is the ability to connect to other components. Connections between components are performed through Ports (discussed below), not through the Resource itself. The use of PortSupplier allows the Resource to return one of any number of Ports that are defined within the component. To provide this functionality, the only interface provided by PortSupplier is getPort( ). As its name implies, getPort returns the port specified in the method's arguments.

ResourceFactory

The SCA provides significant latitude concerning how an instance of a Resource can be created. In its broadest context, a Resource can be created at any time by any entity before it is needed by the framework for a specific waveform. This wide latitude is not necessarily useful because sometimes the framework needs the Resource to be explicitly created. In the cases in which it needs to be explicitly created, the specifications provide the framework with the ResourceFactory class. The ResourceFactory class has only three methods, createResource( ), releaseResource( ), and shutdown( ). The createResource function creates an instance of the desired Resource and returns the reference to the caller, and the release Resource function calls the releaseObject interface in the specified Resource. The shutdown function terminates the ResourceFactory. How the ResourceFactory goes about actually creating the Resource is not described in the specifications. Instead, the specifications provide the developer with the latitude necessary to create the Resource in whichever way seems best for that particular Resource.

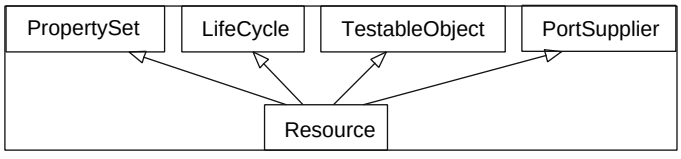


Figure 2.4a-6 Resource class parent structure. The resource inherits a specialized framework functionality.

Port

The Port class is the entry point and, if desired, the exit point of any component. As such, the only function calls that are explicitly listed in the Port definition are connectPort( ) and disconnectPort( ). All other implementation-specific member functions are the actual connections, which, in the most general sense, are guided by the waveform's API. A component can have as many Ports as it requires. The implementation of the Port and the structure that is used to transfer information between the Resource (or the Resource's child) to the Port and back is not described in the specifications and is left up to the developer's discretion. Section 2.4a.3.1 describes the development process and describes a few patterns that can be used to create specific components.

Framework control

The base component classes are the basic building blocks of the waveform, which are relatively simple; the complexity in component development arrives in the implementation of the actual component functionality. When developing or acquiring a framework, the bulk of the complexity is in the framework control classes. These classes are involved in the management of system hardware, systemwide and hardware-specific storage, and deployed waveforms. From a high level, the framework control classes provide all the functionality that one would expect from an OS other than thread and process scheduling.

From Figure 2.4a-4, the framework control classes can be partitioned into three basic pieces: hardware control, waveform control, and system control. The following sections describe each of these pieces in more detail.

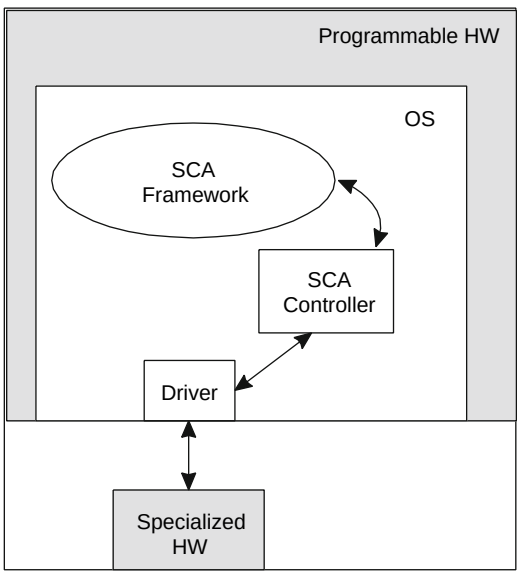


Figure 2.4a-7 Specialized hardware controlled by programmable interface through a proxy structure.

Hardware control

As discussed in Section 2.4a.1, a radio generally comprises a variety of different pieces of hardware. As such, the framework must include the infrastructure necessary to support the variety of hardware that may be used. From a software placement point of view, not all hardware associated with a radio can run software. Thus, the classes that are described to handle hardware can run on the target hardware itself, or it can run in a proxy fashion, as seen in Figure 2.4a-7. When used as a proxy, the hardware control software allows associated specialized hardware to function as if it were any other type of hardware.

The needs of the hardware controller in the SCA are similar to the needs of components supporting a waveform. Furthermore, the component model closely fits the concept of hardware. Thus, the device controllers all inherit from the Resource base class. There are four device controllers: Device, LoadableDevice, ExecutableDevice, and AggregateDevice. Each of these classes is intended to support hardware with increasingly complex needs.

The most fundamental hardware is the hardware that performs some hardwired functionality and that may or may not be configurable. Such hardware has some inherent capacities that may be allocated for specific use. An example of such a piece of hardware is an ADC. The ADC is not programmable, but it is conceivable for an ADC to be tunable, meaning the developer may set the sampling rate or number of quantization bits. In such a case, the available capacity of this tunable hardware depends on whether it is being used for data acquisition of a particular signal or over a particular band. To this end, the only two functions that the Device class includes are allocateCapacity and deallocateCapacity.

Beyond the simple ability to allocate and deallocate capacities, a particular piece of hardware may have the ability to load and unload binary images. These images are not necessarily executable code, but they are images that configure a particular piece of hardware. For example, an FPGA, once loaded with a bit image, is configured as a circuit. The FPGA is never "run," it just acts as a circuit. The LoadableDevice class was created to deal with such hardware, where LoadableDevice inherits from Device. As would be expected, the only two functions that the LoadableDevice class contains are load() and unload().

Finally, there is the more complex hardware that not only has capacities that can be allocated and deallocated as well as memory that can be loaded and unloaded with binary images, but it also can execute a program from the loaded binary image. Such hardware is a GPP or a DSP.

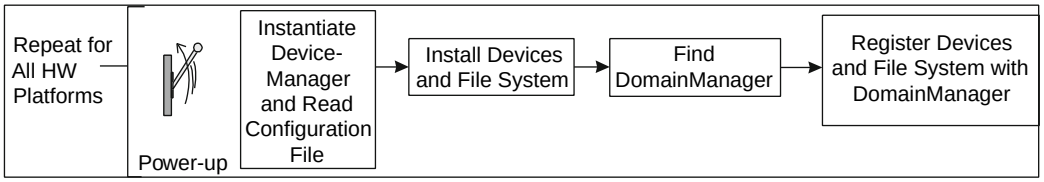


Figure 2.4a-8 Device manager boot-up sequence.

For these types of processors, the SCA uses the ExecutableDevice class. Much like an FPGA, a GPP or DSP has capacities that can be allocated or deallocated (like ports), and memory that can hold binary images, so the ExecutableDevice class inherits from the LoadableDevice class. As would be expected, the two functions that ExecutableDevice supports are execute() and terminate().

DeviceManager

The different hardware controllers behave as stand-alone components, so they need to be created and controlled by another entity. In the case of the SCA, this controller is the DeviceManager. The DeviceManager is the hardware booter; its job is to install all the appropriate hardware for a particular box and to maintain the file structure for that particular set of hardware. The boot-up sequence for the DeviceManager is fairly simple, as shown in Figure 2.4a-8.

As seen in Figure 2.4a-8, when an instance of the DeviceManager is created, it installs all the Devices described in the DeviceManager's appropriate profile and installs whatever file system is appropriate. After installing the hardware and file system, it finds the central controller, in this case the DomainManager, and installs all the devices and file system(s).

In a distributed system with multiple racks of equipment, the DeviceManager can be considered to be the system booter for each separate rack, or each separate board. As a rule, a different DeviceManager is used for each machine that has a different Internet Protocol (IP) address, but that general rule does not have to apply to every possible implementation.

Application control

Two classes in the framework control section of the SCA provide application (waveform) control: Application and ApplicationFactory. ApplicationFactory is the entity that creates waveforms; as such, ApplicationFactory contains a single function, create().

ApplicationFactory

The create() function in the ApplicationFactory is called directly by the user, or the cognition engine in the case of a CR system, to create a specific waveform. The waveform's individual components and their relative connections are described in an eXtensible Markup Language (XML) file, and the component's specific implementation details are described in another XML file. The different XML files that are used to describe a waveform in the SCA are described in the section *Profiles*.

Figure 2.4a-9 shows an outline of the behavior of the ApplicationFactory. As seen in Figure 2.4a-9, the ApplicationFactory receives the request from an external source. Upon receiving this request, it creates an instance of the Application class (discussed next), essentially a handle for the waveform. After the creation of the Application object, the ApplicationFactory allocates the hardware capacities necessary in all the relevant hardware, checks to see if the needed Resources already exist and creates them (through an entity such as the ResourceFactory) when they do not already exist,

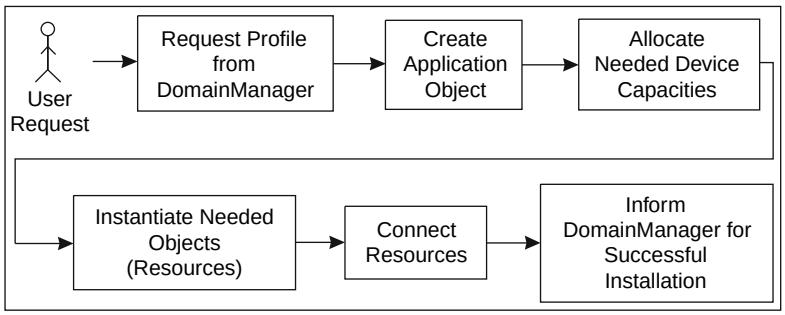


Figure 2.4a-9 Simplified ApplicationFactory create() call for the creation of a waveform.

connects all the Resources, and informs the DomainManager (a master controller discussed in see section *System Control*) that the waveform was successfully created. The other steps in this creation process, such as initializing the Resources, are not included in this description for the sake of brevity. Once the ApplicationFactory has completed the creation process, a waveform comprising a variety of connected components now exists and can be used by the system.

Application class

The ApplicationFactory returns to the caller an instance of the Application class. The Application class is the handle that is used by the environment to keep track of the application. The Application class inherits from the Resource class, and its definition does not add any more function calls, just application identifiers and descriptors. The instance of the Application class that is returned by the ApplicationFactory is the object that the user would call start, stop, and releaseObject on to start, stop, and terminate the application, respectively.

System control

In order for the radio system to behave as a single system, a unifying point is necessary. The specific nature of the unifying point can be as simple as a central registry, or as sophisticated as an intelligent controller. In the SCA, this unifying point is neither. The DomainManager, which is the focal point for the radio, is such an entity, and its task is as a central registry of applications, hardware, and capabilities. Beyond this registry operation, the DomainManager also serves as a mount point for all the different pieces of the hardware's system file. The DomainManager also maintains the central update channels, keeping track of changes in status for hardware and software and also informing the different system components of changes in system status.

The point at which the DomainManager is created can be considered to be the system boot-up. Figure 2.4a-10 shows this sequence. As seen in Figure 2.4a-10, the DomainManager first reads its own configuration file. After determining the different operating parameters, the DomainManager creates the central file system, called a FileManager in the context of the SCA, reestablishes the previous configuration before shutdown, and waits for incoming requests. These requests can be DeviceManagers associating with the DomainManager, or new Applications launched by the Application Factory.

Putting it together

In general, the boot-up sequence of an SCA radio is partitioned into two different sets of steps, Domain boot-up, and one or more Device boot-ups. Once the platform has been installed (i.e., file system(s) and device(s)), the system is ready to accept requests for new waveforms. These requests can arrive either from the user or from some other entity, even a cognitive engine. Figure 2.4a-11 shows a simplified boot-up sequence for the different parts of the SCA radio.

Profiles

The SCA is composed of a variety of profiles to describe the different aspects of the system. The collection of all files describes, "the identity, capabilities, properties, inter-dependencies, and location of the hardware devices and software components that make up the system" [11] and is referred to as the Domain Profile. The Domain Profile is in XML, a language similar to the HyperText Markup Language (HTML), which is used to associate values and other related information to tags. XML is relatively easy for a human to follow and, at the same time, easy for a machine to process.

The relationships among the different profiles is shown in Figure 2.4a-12. As seen in this figure, there are seven file types: Device Configuration Descriptor (DCD), DomainManager Configuration Descriptor (DMD), Software Assembly Descriptor (SAD), Software Package Descriptor (SPD), Device Package Descriptor (DPD), Software Component Descriptor (SCD), and Properties Descriptor (PRF). In addition, there is Profile Descriptor, a file containing a reference to a DCD, DMD, SAD, or SPD.

The profiles have a hierarchical structure that is split into three main tracks, all of which follow a similar pattern. In this pattern, there is an initial file for the track. This initial file contains information about that particular track as well as the names of other files describing other aspects of the track. These other files will

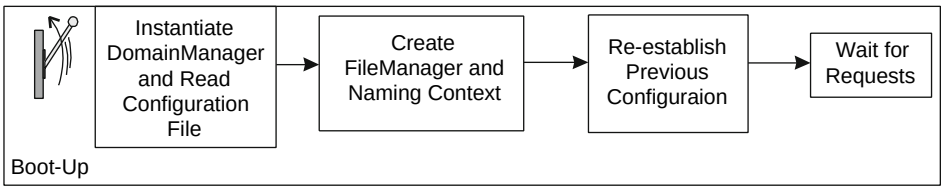


Figure 2.4a-10 DomainManager simplified boot-up sequence.

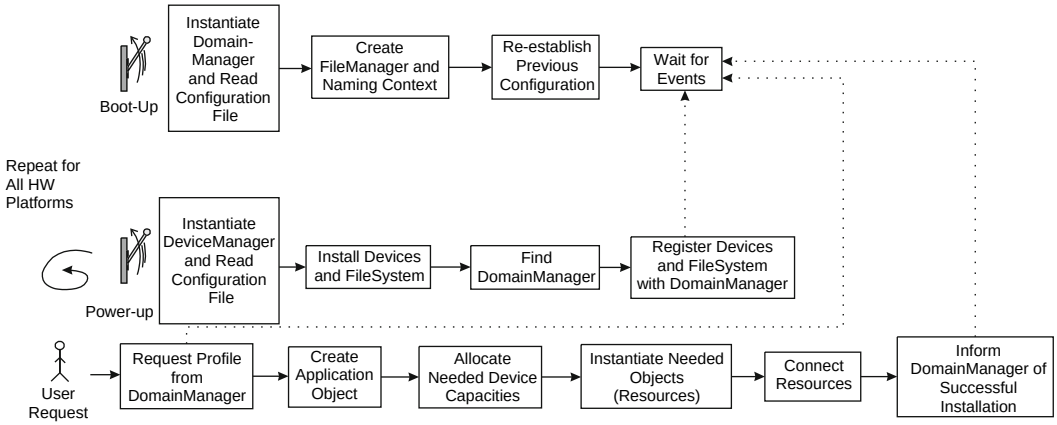


Figure 2.4a-11 Simplified boot-up sequence and waveform creation for SCA radio.

describe their aspect of the system and, when appropriate, will contain a reference to another file, and so on. Three files point to the beginning of a track: DCD, DMD, and SAD. The DCD is the first file that is read by the DeviceManager on boot-up, the DMD is the first file that is read by the DomainManager on boot-up, and the SAD is the first file that is read by the ApplicationFactory when asked to create an application. Each of these files will contain not only information about that specific component, but one or more (in the case of the DMD, only one) references to an SPD. One SPD in this list of references contains information about that specific component. Any other SPD reference is for an additional component in that system: in the case of the DCD, a proxy for hardware; in the case of the SAD, another component in the waveform.

The SPD contains information about that specific component, such as implementation choices, and a link to both a PRF, which contains a list of properties for that component, and the SCD, which contains a list of the interfaces supported by that component. In the case of the DCD, an additional track exists to describe the actual hardware, the DPD. The DPD, much like the SPD, contains information about that specific component, in this case, the hardware side. The DPD, also mirroring the SPD, contains a reference to a PRF, which contains properties information concerning capacities for that particular piece of hardware.

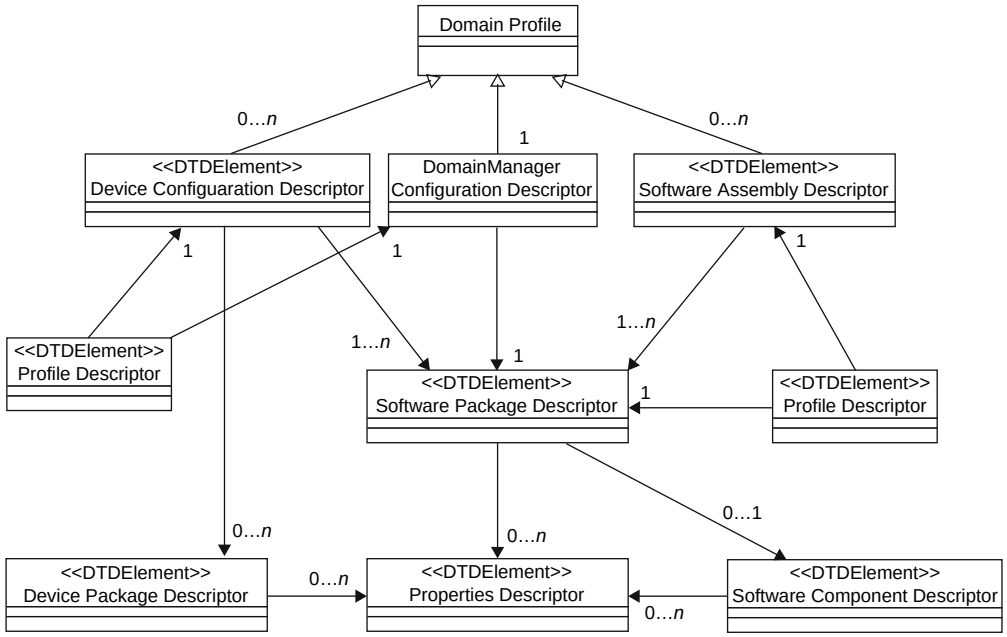


Figure 2.4a-12 Files describing different aspects of an SCA radio.

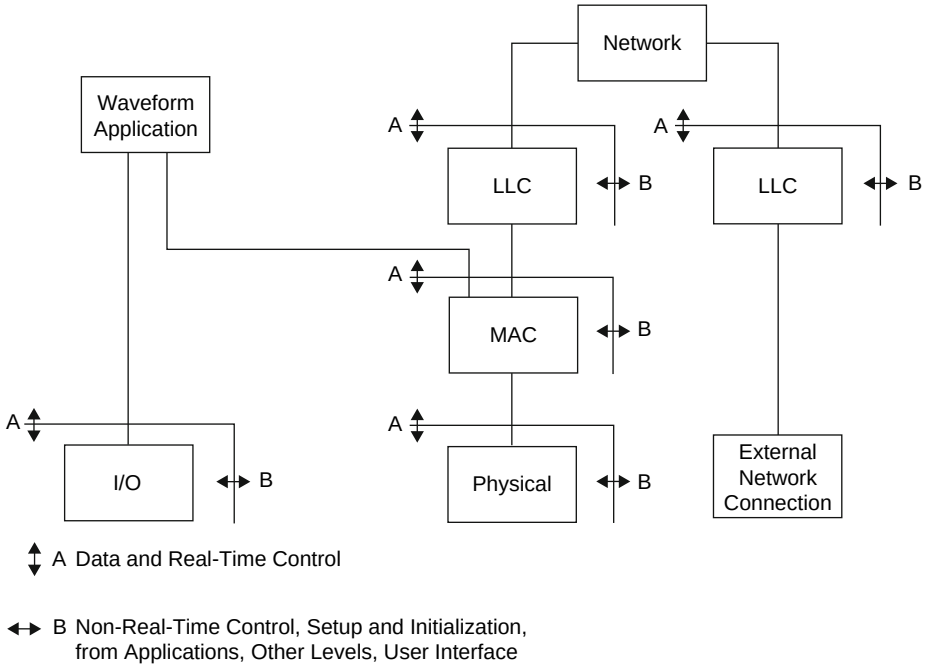


Figure 2.4a-13 SCA API for interlayer communications (I/O = input/output; MAC = medium access control; LLC = logical link control).

Application programming interface

To increase the compatibility between components, it is important to standardize the interfaces; this means standardizing the function names, variable names, and variable types, as well as the functionality associated with each component. The SCA contains specifications for an API to achieve this compatibility goal. The structure of the SCA's API is based on building blocks, where the interfaces are designed in a sufficiently generic fashion such that they can be reused or combined to make more sophisticated interfaces.

There are two types of APIs in the SCA: real-time, shown as A in Figure 2.4a-13, and non-real-time, shown as B in Figure 2.4a-13. Real-time information is both data and time-sensitive control, whereas non-real-time is control information, such as setup and configuration, which does not have the sensitivity of real-time functionality. An interesting aspect of the SCA, API is that it describes interaction between different layers as would be described in the Open Systems Interconnection (OSI) protocol stack.

The API does not include interfacing information for intralayer communications. This means that in an implementation, the communications between two baseband processing blocks, such as a filter and an energy detector, would be outside the scope of the specifications. This limitation provides a limit to the level of granularity that the SCA supports, at least at the level of the API. To increase the granularity to a level comparable to that shown in the example in Section 2.4a.2, an additional API is required.

2.4a.5 Applications

2.4a.5.1 Application software

The application level of the system is where the cognitive engine is likely to operate. At this level, the cognitive engine is just another application in the system. The biggest difference between a cognitive application and a traditional application such as a Web browser is that the cognitive engine has some deep connections with the underlying baseband software. Unlike baseband systems, which are generally light and mainly comprise highly optimized software, application-level design has more leeway in design overhead, allowing the use of a rich and varied set of solutions that sometimes add significant performance overhead. Beyond the typical OSs, such as products by Microsoft or products of the Linux family, and the typical toolkits, such as wxWindows for graphing, or OS-specific solutions such as Microsoft's DCOM, there exist three technologies that can have a significant impact on the application environment for CR: Java, Binary Runtime Environment for Wireless (BREW), and Python.

Java

Java is an object-oriented language developed at Sun Microsystems that allows the development of platform-independent software while also providing some significant housekeeping capabilities to the developer.

Fundamentally, Java is a compiled language that is interpreted in real-time by a virtual machine (VM). The

developer creates the Java code and compiles it into bytecodes by the Java compiler, and the bytecodes are then interpreted by the Java VM (JVM) running on the host processor. The JVM has been developed in some language and compiled to run (as a regular application) on the native processor, and hence should be reasonably efficient when it runs on the target processor. The benefit from such an approach is that it provides the developer the ability to "write once, run anywhere."

Java also provides some housekeeping benefits that can best be defined as dreamlike for an embedded programmer, the most prominent of which is *garbage collection.* Garbage collection is the JVM's ability to detect when memory is no longer needed by its application and to deallocate that memory. This way, Java guarantees that the program will have no memory leaks. In C or C++ development, especially using distributed software such as CORBA, memory management is a challenge, and memory leaks are sometimes difficult to find and resolve. Beyond memory management, Java also has an extensive set of security features that provide limits beyond those created by the developer. For example, Java can limit the program's ability to access native calls in the system, reducing the likelihood that malicious code will cause system problems.

Java has several editions, each tailored to a particular type of environment:

- Java 2 Standard Edition (J2SE) is designed to run in desktop environment, servers, and embedded systems. The embedded design of J2SE relates to reduced overhead inherent to Java, such as reduced memory or reduced computing power.
- Java 2 Enterprise Edition (J2EE) is designed for large, multitier applications as well as for the development of and interactions with Web services.
- Java 2 Micro Edition (J2ME) is designed specifically for embedded and mobile applications. J2ME includes features that are useful in a wireless environment, such as built-in protocols and more robust security features.

Beyond these versions of Java, other features are available that can be useful for the radio developer. For example, under J2ME, there is a JavaPhone API. This API provides application-level support for telephony control, datagram messages (unreliable service over IP), power management, application installation, user profile, and address book and calendar. The JavaPhone API coupled with existing software components can reduce the development time associated with an application suite.

Java has been largely successful, especially in the infrastructure market. In such an environment, any overhead that may be incurred by the use of a real-time interpreter is overwhelmed by the benefit of using Java. For the embedded environment, the success of Java has been limited. There are three key problems with Java: memory, processing overhead, and predictability.

- *Memory:* Most efforts to date in Java for the embedded world revolve around the reduction of the overall memory footprint. One such example is the Connected Limited Device Configuration (CLDC). The CLDC provides a set of API and a set of VM features for such constrained environments.
- *Processing Overhead:* Common sense states that processing overhead is a problem whose relevance is likely to decrease as semiconductor manufacturing technology improves. One of the interesting aspects of processing overhead is that if the minimum supported system data rates for communications systems grow faster than processing power per milliwatt, then processing overhead will become a more pressing problem. The advent of truly flexible software-defined systems will bear out this issue and expose whether it is in fact a problem.
- *Predictability:* Predictability is a real-time concern that is key if the application software is expected to support such functionality as link control. Given that the executed code is interpreted by the JVM, asynchronous events such as memory management can mean that the execution time between arbitrary points of execution in the program can vary. A delay in execution can be acceptable because the design may be able to tolerate such changes. However, large variations in this execution time can cause such problems as dropped calls. A Real-time Extension Specification for Java exists that is meant to address this problem.

Java is a promising language that has met success in several arenas. Efforts within the Java community to bring Java into the wireless world are ongoing and are likely to provide a constantly improving product.

BREW

Qualcomm created the BREW for its CDMA phones. However, BREW is not constrained to just Qualcomm phones. Instead, BREW is an environment designed to run on any OS that simplifies development of wireless applications. BREW supports C/C++ and Java, so it is also largely language independent.

BREW provides a functionality set that allows the development of graphics and other application-level features. BREW is designed for the embedded environment, so the overall footprint and system requirements are necessarily small. Furthermore, it is designed for the management of binaries, so it can download and maintain a series of programs. This set of features is tailored for the handset environment, where a user may download

games and other applications to execute on the handset but lacks the resources for more traditional storage and management structures.

Python

Python is an "interpreted, interactive, OOP language" [12]. Python is not an embedded language and it is not designed for the mobile environment, but it can play a significant role in wireless applications, as evidenced by its use by GNURadio. What makes Python powerful is that it combines the benefits of object-oriented design with the ease of an interpreted language.

With an interpreted language, it is relatively easy to create simple programs that will support some basic functionality. Python goes beyond most interpreted languages by adding the ability to interact with other system libraries. For example, by using Python one can easily write a windowed application through the use of wxWidgets. Just as Python can interact with wxWidgets, it can interact with modules written in C/C++, making it readily extendable.

Python also provides memory management structures, especially for strings, that make application development simpler. Finally, because Python is interpreted, it is OS independent. The system requires the addition of an interpreter to function; of course, if external dependencies exist in the Python software for features such as graphics, the appropriate library also has to exist in the new system.

2.4a.6 Development

Waveform development for sophisticated systems is a complex, multidesigner effort and is well beyond the scope of this book. However, what can be approached is the design of a simple waveform. In this case, the waveform is designed using the SCA, bringing several of the concepts described thus far into a more concrete example. Before constructing a waveform, it is important to build a component.

2.4a.6.1 Component development

Several structures are possible for a component because it is defined only in terms of its interfaces and functionality. At the most basic level, it is possible to create a component class that inherits from both Resource and Port. In such a case, the component would function as a single thread and would be able to respond to only one event at a time. A diagram of this simple component is seen in Figure 2.4a-14.

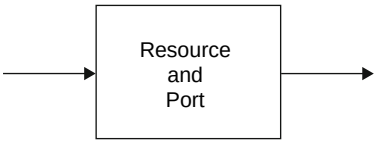


Figure 2.4a-14 Simple resource and port component. The component is both a resource and a port.

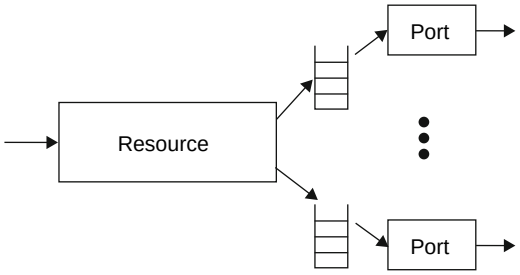


Figure 2.4a-15 Fan-out structure for a component, the resource uses output ports but is the only entry point into the component.

A more sophisticated approach is to separate the output ports into separate threads, each interfacing with the primary Resource through some queue. This approach allows the creation of fan-out structures while at the same time maintaining a relatively simple request response structure. A diagram of this fan-out structure is seen in Figure 2.4a-15.

One of the problems with the component structure shown in Figure 2.4a-15 is that it does not allow for input interfaces that have the same interface name; this limitation increases the difficulty in using the API described in the specifications. To resolve this problem, a fan-in structure needs to be added to the system, even though this creates another level of complexity to the implementation. A way to implement this fan-in structure is to mimic the fan-out structure, and to place each input Port as a separate thread with a data queue separating each of these threads with the functional Resource. An example of this implementation is shown in Figure 2.4a-16.

The approaches shown in Figures 2.4a-14, 2.4a-15, and 2.4a-16 each present a different structure for the same concept. The specific implementation decision for a component is up to the developer, and can be tailored to the specific implementation. Because the components are described in terms of interfaces and functionality, it is possible to mix and match the different structures, allowing the developer even more flexibility.

2.4a.6.2 Waveform development

As an example of the waveform development, assume that a waveform is to be created that splits processing into three parts: baseband processing (assigned to a DSP), link processing (assigned to a GPP), and a user

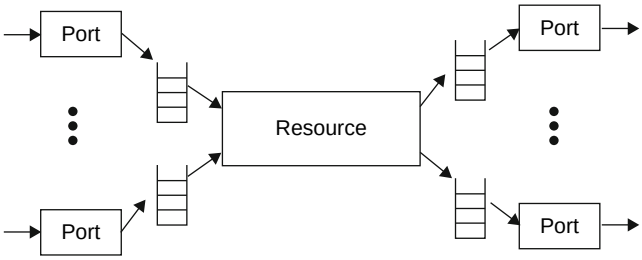


Figure 2.4a-16 Fan-in, fan-out structure for a component. The resource uses both input and output ports.

interface (assigned to the same GPP). A diagram of this waveform is shown in Figure 2.4a-17.

The application shown in Figure 2.4a-17 has several pieces that should be readily recognized; the application was generated by the framework as a result of the ApplicationFactory. Furthermore, the three proxies representing the relevant processing for the radio are also shown. Since the baseband processing is performed on a DSP, the component shown is a proxy for data transfer to and from the DSP. Link processing and the user interface, however, are actually implemented on the GPP. The assembly controller shown is part of the SCA, but not as a separate class. The assembly controller provides control information to all the different deployed components. Because the framework is generic, application-specific information, such as which components to tell to start and stop, must somehow be included. In the case of the SCA, this information is included in the assembly controller, a custom component whose sole job is to interface the application object to the rest of the waveform.

The waveform described in Figure 2.4a-17 does not describe real-time constraints; it assumes that the real-time requirements of the system are met through some other means. This missing piece is an aspect of the SCA that is incomplete.

2.4a.7 Cognitive waveform development

In the context of SDR, a cognitive engine is just another component of a waveform, so the issue in cognitive engine deployment becomes one of component complexity. In the simplest case, a developer can choose to create components that are very flexible, an example of which is seen in Figure 2.4a-18. The flexible baseband processor seen in this figure can respond to requests from the cognitive engine and alter its functionality. A similar level of functionality is also available in the link-processing component. The system shown in Figure 2.4a-18 is a slight modification of the system shown in Figure 2.4a-17. The principal difference in this system is the addition of a reverse-direction set of ports and changed functionality for each component.

The principal problem with the structure shown in Figure 2.4a-18 is that it places all the complexity of the system onto each separate component. Furthermore, the tie-in between the cognitive engine and the rest of the waveform risks the engine implementation to be limited to this specific system. An alternate structure is to create a whole waveform for which the only functionality is a cognitive engine, as seen in Figure 2.4a-19.

The cognitive engine waveform shown in Figure 2.4a-19 has no link to a waveform. To create this link, a link is created between the ApplicationFactory and the cognitive engine waveform. The cognitive engine can then request that the ApplicationFactory launch new waveforms. These new waveforms perform the actual communications work, which is evaluated by the functioning cognitive engine. If performance on the new waveform's communications link falls within some determined parameter, the cognitive engine can terminate the existing communications link waveform and request a new waveform (with different operating parameters or

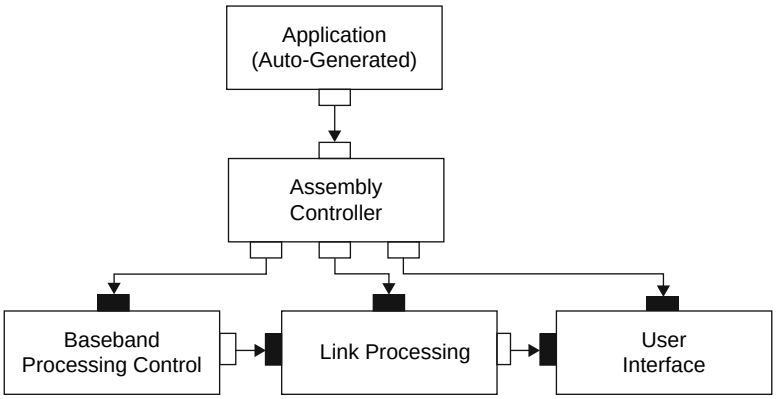


Figure 2.4a-17 Simple SCA application. This example is made up of three functional components and one control component.

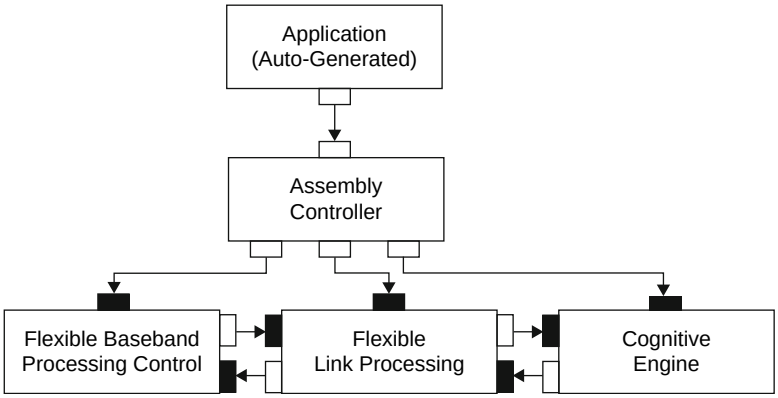


Figure 2.4a-18 Simple cognitive waveform. The waveform performs both communications and cognitive functionality.

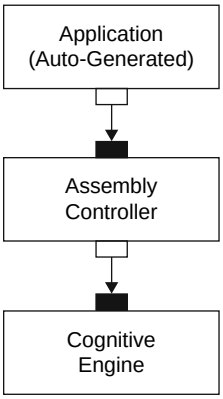


Figure 2.4a-19 Cognitive engine waveform. The waveform performs only cognitive functionality. It assumes communications functionality is performed by other waveforms.

a different structure altogether) from the ApplicationFactory. This structure is seen in Figure 2.4a-20.

An aspect of Figure 2.4a-20 that is apparent is that the Port structure evident at the waveform level, as seen in Figure 2.4a-19, scales up to interwaveform communications. Another aspect of this approach is that the cognitive waveform does not have to be collocated with the communications waveform. As long as timing constraints are met, the cognitive waveform can be placed anywhere within the network that has access to the system. This aspect of the deployment of the waveforms allows the concept of a CR within the SCA to easily extend to a cognitive network, where a single cognitive engine can control multiple flexible radios, as seen in Figure 2.4a-21.

With SDR and an astute selection of processing and RF hardware on the part of a developer, it is possible to create highly sophisticated systems that can operate well at a variety of scales, from simple single-chip mobile devices all the way up to multitiered cognitive networks.

2.4a.8 Summary

Although an SDR is not a necessary building block of a CR, the use of SDR in CR can provide significant capabilities to the final system. An SDR implementation is a system decision, in which the selection of both the underlying hardware composition and the software architecture are critical design aspects.

The selection of hardware composition for an SDR implementation requires an evaluation of a variety of

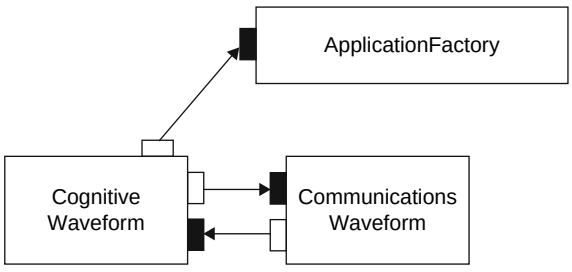


Figure 2.4a-20 Multi-waveform cognitive support. The stand-alone cognitive waveform requests new waveforms from the SCA ApplicationFactory.

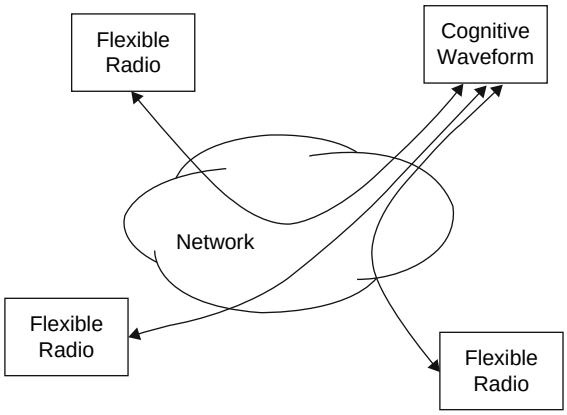


Figure 2.4a-21 SCA-enabled cognitive network composed of multiple cognitive nodes.

aspects, from the hardware's ability to support the required signals to other performance aspects, such as power consumption and silicon area. Traditional approaches can be used to estimate the needs at the RF and data acquisition levels. At the processing stage, it is possible to create an estimate of a processing platform's ability to be able to support a particular set of signal processing functions. With such an analysis, it is possible to establish the appropriate mix of general-purpose processors, DSPs, FPGAs, and CCMs for a particular set of signal processing needs.

In order to mimic the nature of a hardware-based radio, with components such as mixers and amplifiers, CBP is a natural way to consider software for SDR. In CBP, components are defined in terms of their interfaces and functionality. This definition provides the developer with significant freedom on the specific structure of that particular component.

Even though a developer may choose to use CBP for the design of an SDR system, a substantial infrastructure is still needed to support SDR implementations. This infrastructure must provide basic services, such as the creation and destruction of waveforms, as well as general system integration and maintenance. The goal of a software architecture is to provide this underlying infrastructure. The SCA is one such architecture. The SCA provides the means to create and destroy waveforms, manage hardware and distributed file systems, and manage the configuration of specific components.

Finally, beyond programming methodologies and architectures are the actual languages that one can use for development of waveforms and the specific patterns that are chosen for the developed software. The various languages have different strengths and weaknesses. C++ and Java are the dominant languages in SDR today. Python, a scripting language, has become increasingly popular in SDR applications, and is likely to be an integral part of future SDR development.

Much like the language selection, a design pattern for a particular component can have a dramatic effect on the capabilities of the final product. Design patterns that focus on flexibility can be more readily applied to cognitive designs, from the most basic node development all the way up to full cognitive networks.

References

1. J. Neel, J.H. Reed and M. Robert, "A Formal Methodology for Estimating the Feasible Processor Solution Space for a Software Radio," in *Proceedings of the 2005 Software Defined Radio Technical Conference and Product Exposition*, November 2005, Orange Co, CA.
2. T. Budd, *An Introduction to Object-Oriented Programming, Third Edition*, Addison-Wesley, 2001, Boston, MA.
3. M. Weisfeld, *The Object-Oriented Thought Process, Second Edition*, Sams, 2003,Indianapolis, IN.
4. E. Gamma, R. Helm, R. Johnson and J. Vlissides, *Design Patterns: Elements of Reusable Object-Oriented Software*, Addison-Wesley, 1995, Boston, MA.
5. A. Shalloway and J.R. Trott, *Design Patterns Explained: A New Perspective on Object-Oriented Design, Second Edition*, Addison-Wesley, 2005, Boston, MA.
6. J. Kerievsky, *Refatcoring to Patterns*, Addison-Wesley, 2005, Boston, MA.
7. http://www.gnu.org/software/gnuradio/
8. http://ossie.mprg.org/
9. http://www.crc.ca/en/html/crc/home/research/satcom/rars/sdr/sdr
10. M. Henning and S. Vinoski, *Advanced CORBA Programming with* C++, Addison-Wesley, 1999, Boston, MA.
11. JTRS Joint Program Office, JTRS-5000, "Software Communications Architecture Specification, SCA V3. 0," August 2004, San Deigo, CA.
12. http://www.python.org/

Chapter 2.4b

2.4b

Cognitive radio: The technologies required

Pablo Robert and Bruce Fette

2.4b.1 Introduction

Technology is never adopted for technology's sake. For example, only hobbyists used personal computers (PCs) until a spreadsheet program, a "killer application," was developed. Then business needs and the benefits of small computers became apparent and drove PC technology into ubiquitous use. This led to the development of more applications, such as word processors, e-mail, and more recently the World Wide Web (WWW). Similar development is under way for wireless communication devices.

Reliable cellular telephony technology is now in widespread use, and new applications are driving the industry. Where these applications go next is of paramount importance for product developers. Cognitive radio (CR) is the name adopted to refer to technologies believed to enable some of the next major wireless applications. Processing resources and other critical enabling technologies for wireless killer applications are now available.

This chapter presents a CR roadmap, including a discussion of CR technologies and applications. Section 2.4b.2 presents a taxonomy of radio maturity, and Sections 2.4b.3 and 2.4b.4 present more detailed discussions. Sections 2.4b.5 and 2.4b.6 are about enabling and required technologies for CRs. They present three classes of cognitive applications, one of which may be the next killer application for wireless devices. Conjectures regarding the development of CR are included in Section 2.4b.7 with arguments for their validity. Highlights of this chapter are discussed in the summary in Section 2.4b.8, which emphasizes that the technologies required for CR are presently available.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2006 Elsevier Inc. All rights reserved

2.4b.2 Radio flexibility and capability

More than 40 different types of military radios (not counting variants) are currently in operation. These radios have diverse characteristics; therefore, a large number of examples can be drawn from the pool of military radios. This section presents the continuum of radio technology leading to the software-defined radio (SDR). Section 2.4b.3 continues the continuum through to CR.

The first radios deployed in large numbers were "single-purpose" solutions. They were capable of one type of communication (analog voice). Analog voice communication is not particularly efficient for communicating information, so data radios became desirable, and a generation of data-only radios was developed. At this point, our discussion of software and radio systems begins. The fixed-point solutions have been replaced with higher data rate and voice capable radios with varying degrees of software integration. This design change has enabled interoperability, upgradability, and portability.

It will be clear from the long description of radios that follows that there have been many additional capabilities realized in radios over their long history. SDRs and even more advanced systems have the most capabilities, and additional functions are likely to be added over time.

2.4b.2.1 Continuum of radio flexibility and capability

Basing CR on an SDR platform is not a requirement, but it is a practical approach at this time because SDR flexibility allows developers to modify existing systems with

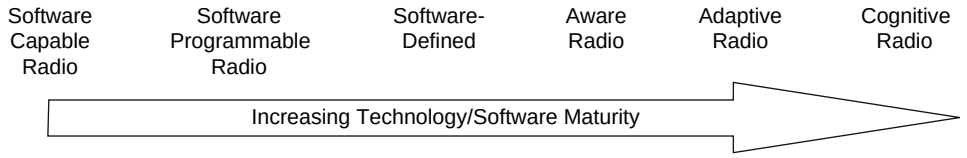


Figure 2.4b-1 SDR technology continuum. As software sophistication increases, the radio system capabilities can evolve to accommodate a much broader range of awareness, adaptivity, and even the ability to learn.

little or no new hardware development, as well as to add cognitive capabilities. The distinction of being a CR is bestowed when the level of software sophistication has risen sufficiently to warrant this more colorful description. Certain behaviors, discussed in this chapter, are needed for a radio to be considered a CR.

Historically, radios have been fixed-point designs. As upgrades were desired to increase capability, reduce life cycle costs, and so forth, software was added to the system designs for increased flexibility. In 2000, the Federal Communications Commission (FCC) adopted the following definition for software radios: "A communications device whose attributes and capabilities are developed and/or implemented in software" [1]. The culmination of this additional flexibility is an SDR system, as software capable radios transitioned into software programmable radios and finally became SDRs. The next step along this path will yield aware radios, adaptive radios, and finally CRs (see Figures 2.4b-1 and 2.4b-2).

2.4b.2.2 Examples of software capable radios

Several examples of software capable radios are shown in Figure 2.4b-2 and detailed in Table 2.4b-1. The common characteristics of these radios are fixed modulation capabilities, relatively small number of frequencies, limited data and data rate capabilities, and finally the ability to handle data under software control.

2.4b.2.3 Examples of software programmable radios

Several examples of software programmable radios are shown in Figure 2.4b-2 and detailed in Table 2.4b-2. The common characteristics of these radios are their ability to add new functionality through software changes and their advanced networking capability.

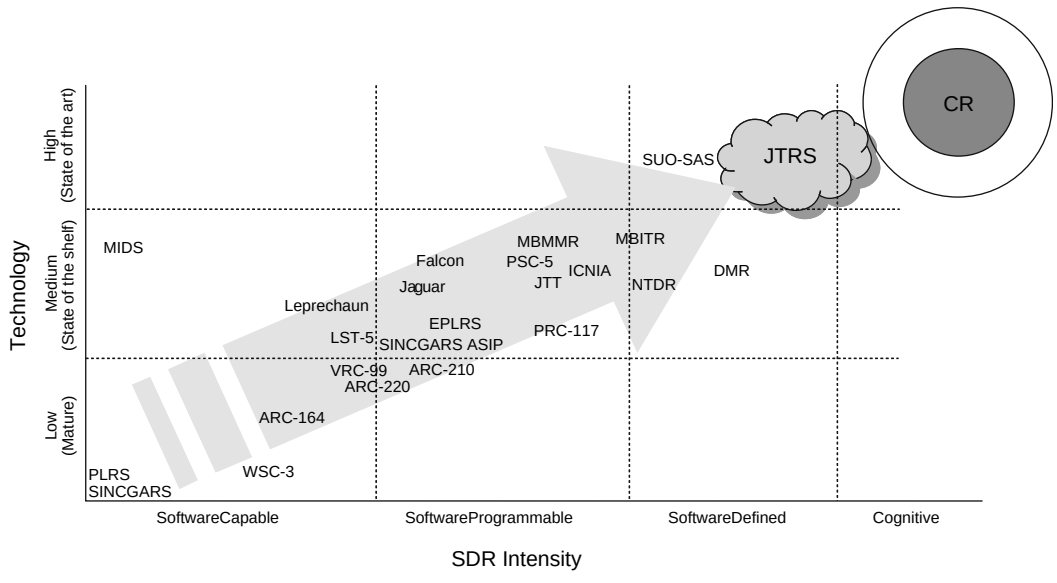


Figure 2.4b-2 Examples of software radios. The most sophisticated software control and applications are reaching cognitive levels (see Tables 2.4b-1–2.4b-3 for descriptions).[1]

[1] Note that the radios that are part of Figure 2.4b-2 and Tables 2.4b-1–2.4b-3 are based on the best available public information and are intended only to notionally indicate a distribution of some of the well-known radios. For additional, accurate information about capabilities, contact the respective manufacturers.

Table 2.4b-1 Selected software capable radios.

SINCGARS (Single-Channel Ground and Airborne Radio System)	A family of very high frequency—frequency modulation (VHF-FM) radios that provides a primary means of command and control. SINCGARS has frequency-hopping capability, and certain US Air Force versions operate in other bands using amplitude modulation (AM) waveforms. The SINCGARS family of radios has the capability to transmit and receive voice and tactical data, and record traffic on any of 2320 channels (25 kHz) between 30 and 88 MHz. A SINCGARS with an Internet controller is software capable [2].
PLRS (Position Location Reporting System)	A command-and-control aide that provides real-time, accurate, three-dimensional (3D) positioning, location, and reporting information for tactical commanders. The jam-resistant ultra high-frequency (UHF) radio transceiver network automatically exchanges canned messages that are used to geolocate unit positions to 15-m accuracy. Commanders use PLRS for situational awareness. PLRS employs a computer-controlled, crypto-secured master station and an alternate master station to ensure system survivability and continuity of operations. The network, under master station management, automatically uses user units as relays to achieve over-the-horizon (OTH) transmission and to overcome close-in terrain obstructions to line-of-sight (LOS) communications. When a rugged portable computer (PC) is used with the user unit, it becomes a mini-command-and-control station with a local area map superimposed with position and identification (ID) information. The computer interface and network control make PLRS a software capable radio system [3].
AN/WSC-3	A Demand Assigned Multiple Access (DAMA) satellite communication (SATCOM) terminal. It meets tight size and weight integration requirements. This single-waveform radio has a computer (Ethernet) interface and is software capable [4].
AN/ARC-164 HaveQuick II	A LOS UHF-AM radio used for air-to-air, air-to-ground, and ground-to-air communications. ARC-164 radios are deployed on all US Army rotary wing aircraft and provide anti-jam, secure communications links for joint task force missions in the tactical air operations band. This radio operates as a single-channel (25 kHz) or a frequency-hopping radio in the 225–399.975 MHz band. The aircraft radio transmits at 10 W output power and can receive secure voice or data. The ARC-164 data-handling capability makes it software capable. It has an embedded electronic counter-countermeasure (ECCM) anti-jam capability. The 243.000 MHz guard channel can be monitored. One model of the AN/ARC-164 is interoperable with SINCGARS [5].
AN/ARC-220	The standard high-frequency (HF) (2.0–29.999 MHz) radio for US Army aviation. It includes secure voice and data communications on any of 280,000 frequencies. The system uses software-realized DSP. Using MIL-STD-2217, software updates can be made over a MIL-STD-1553B bus. ARC-220 data processing capabilities optionally include e-mail applications. The ARC-220 is software capable. MIL-STD-148–141A automatic link establishment (ALE) protocols improve connectivity on behalf of ARC-220 users. ALE executes in microprocessors and is particularly interesting as a cognitive application because of its use of a database and its sounding of the RF environment for sensing. GPS units may be interfaced with the radio, providing geolocation information [6].
AN/VRC-99	A secure digital network radio used for high data rate applications. This 1.2–2.0 GHz broadband direct-sequence spread-spectrum radio has National Security Agency (NSA) certified high-assurance capabilities and provides users with 31 RF channels. It supports TDMA and FDMA (frequency division multiple access). The networking capabilities are software programmable. AN/VRC-99 provides digital battlespace users with the bandwidth to support multimedia digital terminal equipment (DTE) of either an army command-and-control vehicle (C2V), army battle-command vehicle (BCV), or a marine corps advanced amphibious assault vehicle (A3V) with a single wideband secure waveform that supports voice, workstation, data, and imagery with growth for video requirements [7].

(Continued)

Table 2.4b-1 Selected software capable radios.—Cont'd

LST-5E	A software capable UHF tactical SATCOM/LOS transceiver with embedded communications security (COMSEC). The LST-5E provides the user with a single unit for high-grade half-duplex secure voice and data over both wideband (25 kHz) AM/FM and narrowband (5 kHz) 1200 bits per second (bps) binary phase shift keying (BPSK) and 2400 bps BPSK. Current applications for the LST-5E include manpack, vehicular, airborne, shipborne, remote, and fixed stations. The terminal is compatible with other AM or FM radios that operate in the 225–399.995 MHz frequency band [8].
AN/PRC-6725 or Leprechaun	A handheld or wearable tactical radio. It can be programmed from a frequency fill device, laptop or PC, system base station, or it can be cloned from another radio [9].
MBITR (MultiBand Intra/Inter Team Radio)	Provides voice communications between infantry soldiers. The Land Warrior Squad Radio is a SINCGARS-compatible, eight-channel radio. MBITR is a software capable design based on the commercially available PRC-6745 Leprechaun radio [10].
CSEL (Combat Survivor/Evader Locator)	Provides UHF communications and location for the purposes of Joint Search and Rescue Center operations. CSEL uses GPS receivers for geolocation. Two-way, OTH, secure beaconing through communication satellites allows rescue forces to locate, authenticate, and communicate with survivors/evaders from anywhere in the world. The handheld receivers are one segment of the CSEL command-and-control system. The satellite-relay base station, Joint Search and Rescue Center software suite, and radio set adapter units interface with the UHF SATCOM network to provide CSEL capability. Upgrades through software loads make CSEL radios software capable [11].
MIDS (Multifunction Information Distribution System)	A direct-sequence spread-spectrum, frequency-hopping, anti-jam radio that supports the Link 16 protocol for communication between aircraft. Operates in the band around 1 GHz. Originally called Joint Tactical Information Distribution System (JTIDS), this radio waveform has been redeveloped on a new hardware platform, and is now being converted to a JTRS-compliant radio. When this conversion is completed it will be an SDR. It is currently in various stages of development and production as an interoperable waveform for Europe and the United States [12].

2.4b.2.4 Examples of SDR

Only a few fully SDR systems are available, as shown in Figure 2.4b-2 and detailed in Table 2.4b-3. The common characteristic of SDR systems is complete adjustability through software of all radio operating parameters.

Other products related to SDR include GNURadio and the Vanu Anywave™ Base Station. GNURadio is a free software toolkit that is available on the Internet. It allows anyone to build a narrowband SDR. Using a Linux-based computer, a radio frequency (RF) front-end and an analog-to-digital converter (ADC), one can build a software-defined receiver. By adding a digital-to-analog converter (DAC) and possibly a power amplifier, one can build a software-defined transmitter [25].

The Vanu Anywave™ Base Station is a software-defined system that uses commercial off-the-shelf (COTS) hardware and proprietary software to build a wireless cellular infrastructure. The goal is simultaneous support for multiple standards, reduced operating expenses, scalability, and future proofing (cost-effective migration) [26].

2.4b.3 Aware, adaptive, and CRs

Radios that sense all or part of their environment are considered aware systems. Awareness may drive only a simple protocol decision or may provide network information to maintain a radio's status as aware. A radio must additionally autonomously modify its operating parameters to be considered adaptive. This may be accomplished via a protocol or programmed response. When a radio is aware, adaptive, and learns, it is a CR [27]. Only cutting-edge research and demonstration examples of aware, adaptive, or CRs are available currently.

2.4b.3.1 Aware radios

A voice radio inherently has sensing capabilities in both audio (microphone) and RF (receiver) frequencies. When these sensors are used to gather environmental information, it becomes an aware radio. The local RF spectrum may be sensed in pursuit of channel estimates, interference, or signals of interest. Audio inputs may be

Table 2.4b-2 Selected examples of software programmable radios.

AN/ARC-210	Provides fully digital secure communications in tactical and air traffic control (ATC) environments (30–400 MHz). Additionally, the radio provides 8.33 kHz channel spacing capability to increase the number of available ATC frequencies, and provides for growth to new very-high-frequency (VHF) data link modes. Currently provided functions are realized through software and provide integrated communications systems with adaptability for future requirements with little or no hardware changes. In other words, the ARC-210 is software programmable. The ARC-210 is integrated into aircraft and operated via MIL-STD-1553B data bus interfaces. Remote control is also available for manual operation. This radio is interoperable with SINCGARS (Single-Channel Ground and Airborne Radio System) and HaveQuick radios [13].
Racal 25	A compliant Project 25 public safety radio operating in the 136–174 MHz band. Its 5 W peak transmit power, rugged and submersible housing, and digital voice with Digital Encryption Standard (DES) make it an advanced radio. The radio uses DSP and flash memory architectures, and supports 12 and 16 kbps digital voice and data modes. Racal posts software upgrades on a protected Internet site. Software programmability enables field upgrades and is possible due to its DSP and flash memory-based architecture [14].
SINCGARS ASIP (Advanced System Improvement Program)	Interoperates with SINCGARS radios and enhances operational capability in the Tactical Internet (TI) environment. The ASIP models reduce size and weight, and provide further enhancements to operational capability in the TI environment. SINCGARS ASIP radios are software programmable and provide improved data capability, improved FEC for low-speed data modes, a GPS interface, and an Internet controller that allows them to interface with EPLRS and Battlefield Functional Area host computers. The introduction of Internet Protocol (IP) routing enables numerous software capabilities for radio systems [2].
EPLRS (Enhanced Position Location Reporting System)	An AN/TSQ-158 that transmits digital information in support of tactical operations over a computer-controlled, TDMA communications network. EPLRS provides two major functions: data distribution and position location and reporting. EPLRS uses a frequency-hopping, spread-spectrum waveform in the UHF band. The network architecture is robust and self-healing. Radio firmware may be programmed from external devices. The peak data rate of 525 kbps supports application layer radio-to-radio interaction [15].
AN/PRC-117F	A software programmable, multiband, multimode radio (MBMMR). The PRC-117F operates in the 30–512 MHz band. Embedded COMSEC, SAT-COM (satellite communication), SINCGARS, HaveQuick, and ECCM capabilities are standard. Various software applications, such as file transfer, Transmission Control Protocol (TCP) with IP, and digital voice, are included in this software programmable radio [7].
Jaguar PRC-116	In service in more than 30 nations, including the UK (Army) and US (Navy). It is a frequency-hopping, software programmable radio with considerable ECCM capability, resisting jamming by constantly shifting hopsets to unjammed frequencies. Security is further heightened by use of a scrambler. The Jaguar-V can also be used for data transmission at a rate of 16 kbps, and it may tolerate up to 50 radio nets, each with dozens of radios, at once; if each net is frequency hopping in a different sequence, it will still transmit to all of them [16].
JTT (Joint Tactical Terminal)	A high-performance, software programmable radio. Its modular functionality is backward and forward compatible with the Integrated Broadcast Service (IBS). Using a software download, JTT can accept changes in format and protocol as IBS networks migrate to a common format. Subsequent intelligence terminals require total programmability of frequency, waveform, number of channels, communication security, and transmission security, making subsequent terminals SDRs [17].

(Continued)

Table 2.4b-2 Selected examples of software programmable radios.—Cont'd

NTDR (Near-Term Digital Radio) system	A mobile packet data radio network that links Tactical Operations Centers (TOCs) in a brigade area (up to 400 radios). The NTDR system provides self-organizing, self-healing network capability. Network management terminals provide radio network management. The radios interface with emerging Army Battle Command System (ABCS) automated systems and support large-scale networks in mobile operations with efficient routing software that supports multicast operations. Over-the-air programmability eliminates the need to send maintenance personnel to make frequency changes [18].
AN/PRC-138 Falcon	A manpack or vehicular-mounted HF and VHF radio set. Capabilities include frequency coverage of 1.6–60 MHz in SSB/CW/AME (single-sideband, continuous wave, AM equivalent) and FM in the VHF band and 100 preset channels. In the data mode, the AN/PRC-138 offers a variety of compatible modem waveforms that allows it to be integrated into existing system architectures. Specific features include embedded encryption, ALE, embedded high-performance HF data modems, improved power consumption management, and variable power output [19].
MBMMR (MultiBand, MultiMode Radio)	An AN/PSC-5D(C) that enhances interoperability among special operation forces units. The MBMMR supports LOS and SATCOM voice and data in six basic modes: LOS, Maritime, HaveQuick I/II, SINCGARS, SATCOM, and DAMA. Additional features include embedded TI Range Extension and Mixed Excitation Linear Prediction (MELP) voice coding. Over-the-air rekeying (OTAR), extended 30 to 420 MHz band, MIL-STD-188–181B high data rate in LOS communications and SATCOM, and MIL-STD-188–184 embedded advanced data controller are supported [20].
MBITR (MultiBand Intra/Inter Team Radio)	Designated AN/PRC-148, the MBITR provides AM/FM voice and data communications in the 30–512 MHz band. Development of the MBITR is an outgrowth of Racal's work on DSP and flash memory. MBITR is JTRS SCA 2.2 compliant and does not require a JTRS waiver. The information security (INFOSEC) capabilities are software programmable [21].

used for authentication or context estimates or even natural language understanding or aural human–machine interface (HMI) interactions. Added sensors enable an aware radio to gather other information, such as chemical surroundings, geolocation, time of day, biometric data, or even network quality of service (QoS) measures. The key characteristic that raises a radio to the level of aware is the consolidation of environmental information not required to perform simple communications. Utilization of this information is not required for the radio to be considered aware. There is no communication performance motivation for developing this class of aware radios, and it is expected that this set will be sparse.

One motivation for an aware radio is providing information to the user. As an example, an aware radio may provide a pull-down menu of restaurants within a user-defined radius. The radio may gather this information, in the future, from low-power advertisement transmissions sent by businesses to attract customers. A military application may be a situational awareness body of information that includes a pre-defined set of radios and their relative positions. As an example, the radios exchange global positioning system (GPS) coordinates in the background, and the aware radios gather the information for the user and provide it on request. The radio is not utilizing the information but is aware of the situation.

One example of an aware radio is the code division multiple access (CDMA) based cellular system proposed by Chen et al. [28]. This system is aware of QoS metrics and makes reservations of bandwidth to improve overall QoS. Another example of an aware radio is the orthogonal frequency division multiplexing (OFDM) based energy aware radio link control discussed by Bougard et al. [29].

2.4b.3.2 Adaptive radios

Frequency, instantaneous bandwidth, modulation scheme, error correction coding, channel mitigation strategies such as equalizers or RAKE filters, system timing (e.g., a time division multiple access [TDMA] structure), data rate (baud timing), transmit power, and even filtering characteristics are operating parameters that may be adapted. A frequency-hopped spread-spectrum radio is not considered adaptive because once programmed for a hop sequence, it is not changed. A frequency-hopping radio that changes hop pattern to reduce collisions may be considered adaptive. A radio that supports multiple channel bandwidths is not adaptive, but a radio that

Table 2.4b-3 Selected examples of SDRs.

DMR (Digital Modular Radio)	A full SDR capable of interoperability with tactical systems such as HF, DAMA, HaveQuick, and SINCGARS, as well as data link coverage for Link 4A and Link 11. These systems are programmable and include software-defined cryptographic functions. The US Navy is committed to migrating DMR to SCA compliance to allow the use of JTRS JPO provided waveforms. The DMR may be reconfigured completely via on-site or remote programming over a dedicated LAN or wide area network (WAN). The four full-duplex programmable RF channels with coverage from 2.0 MHz to 2.0 GHz require no change in hardware to change waveforms or security. The system is controlled, either locally or across the network, by a Windows®-based HMI [23].
JTRS (Joint Tactical Radio System)	A set of current radio procurements for fully SDRs. These radio systems are characterized by SCA compliance that specifies an operating environment that promotes waveform portability. The JTRS JPO is procuring more than 30 waveforms that will ultimately be executable on JTRS radio sets. System packaging ranges from embeddable single-channel form factors to vehicle-mounted multi-channel systems. JTRS radios are the current state-of-the-art technology and have the highest level of software sophistication ever embedded into a radio [24].
SUO-SAS (Small Unit Operations—Situational Awareness System)	Developed by DARPA to establish the operational benefits of an integrated suite of advanced communication, navigation, and situation awareness technologies. It served as a mobile communications system for small squads of soldiers operating in restrictive terrain. SUO-SAS provides a navigation function utilizing RF ranging techniques and other sensors to provide very high accuracy [22].

changes instantaneous bandwidth and/or system timing parameters in response to offered network load may be considered adaptive. If a radio modifies intermediate frequency (IF) filter characteristics in response to channel characteristics, it may be considered adaptive. In other words, if a radio makes changes to its operating parameters, such as power level, modulation, frequency, and so on, it may be considered an adaptive radio.

At this time, two wireless products exhibit some degree of adaptation: the digital European cordless telephone (DECT) and 802.11a.

DECT can sense the local noise floor and interference of all the channels from which it may choose. Based on this sensing capability, it chooses to use the carrier frequencies that minimize its total interference. This feature is built into hardware, however, and not learned or software adaptive; thus, DECT is not normally considered an adaptive radio.

802.11 a has the ability to sense the bit error rate (BER) of its link, and to adapt the modulation to a data rate and a corresponding forward error correction (FEC) that set the BER to an acceptably low error rate for data applications. Although this is adaptive modulation, 802.11 implementations generally are dedicated purpose-fixed application-specific integrated circuit (ASIC) chips, not software defined, and thus 802.11 is not normally considered to be an adaptive radio.

2.4b.3.3 Cognitive radios

A CR has the following characteristics: sensors creating awareness of the environment, actuators to interact with the environment, a model of the environment that includes state or memory of observed events, a learning capability that helps to select specific actions or adaptations to reach a performance goal, and some degree of autonomy in action.

Since this level of sophisticated behavior may be a little unpredictable in early deployments, and the consequences of "misbehavior" are high, regulators will want to constrain a CR. The most popular suggestion to date for this constraint is a regulatory policy engine that has machine-readable and interpretable policies.

Machine-readable policy-controlled radios are attractive for several reasons. One feature is the ability to "try out" a policy and assess it for impacts. The deployment may be controlled to a few radios on an experimental basis, so it is possible to assess the observation and measurement of the behaviors. If the result is undesirable, the policies may be removed quickly. This encourages rapid decisions by regulatory organizations. The policy-driven approach is also attractive because spatially variant or even temporally variant regulations may be deployed. As an example, when a radio is used in one country, it is subject to that country's regulations, and when the user carries it

Table 2.4b-4 Properties of advanced radio classes.

Radio property	Software capable radio	Software programmable radio	Software-defined radio	Aware radio	Adaptive radio	Cognitive radio
Frequency hopping	X	X	X	X	X	X
Automatic link establishment (i.e., channel selection)	X	X	X	X	X	X
Programmable crypto	X	X	X	X	X	X
Networking capabilities		X	X	X	X	X
Multiple waveform interoperability		X	X	X	X	X
In-the-field upgradable		X	X	X	X	X
Full software control of all signal processing, crypto, and networking functionality			X	*	*	*
QoS measuring/ channel state information gathering				X	X	X
Modification of radio parameters as function of sensor inputs					X	X
Learning about environment						X
Experimenting with different settings						X

* The industry standards organizations are in the process of determining the details of what properties should be expected of aware, adaptive, and CRs.

to a new country, the policy may be reloaded to comply in the new jurisdiction. Also, if a band is available for use during a certain period but not during another, a machine-readable policy can realize that behavior.

The language being used to describe a CR is based on the assumption of a smart agent model, with the following capabilities:[2]

- Sensors creating awareness in the environment.
- Actuators enabling interaction with the environment.
- Memory and a model of the environment.
- Learning and modeling of specific beneficial adaptations.
- Specific performance goals.
- Autonomy.
- Constraint by policy and use of inference engine to make policy-constrained decisions.

The first examples of CRs were modeled in the Defense Advanced Research Projects Agency (DARPA) NeXt Generation (XG) radio development program. These radios sense the spectrum environment, identify an unoccupied portion, rendezvous multiple radios in the unoccupied band, communicate in that band, and vacate the band if a legacy signal re-enters that band. These behaviors are modified as the radio system learns more about the environment; the radios are constrained by regulatory policies that are machine interpretable. The first demonstrations of these systems took place late in 2004 and in 2005 [30].

[2] A smart agent also has the ability to *not* use some or all of the listed capabilities.

2.4b.4 Comparison of radio capabilities and properties

Table 2.4b-4 summarizes the properties for the classes of advanced radios described in the previous sections. Classes of radios have "fuzzy boundaries," and the comparison shown in the table is broad. There are certain examples of radios that fall outside the suggestions in the table. A CR may demonstrate most of the properties shown, but is not required to be absolutely reparameterizable. Note that the industry consensus is that a CR is not required to be an SDR, even though it may demonstrate most of the properties of an SDR. However, there is also consensus that the most likely path for development of CRs is through enabling SDR technology.

2.4b.5 Available technologies for CRs

The increased availability of SDR platforms is spurring developments in CR. The necessary characteristics of an SDR required to implement a practical CR are excess computing resources, controllability of the system operating parameters, affordability, and usable software development environments including standardized application programming interfaces (APIs). This section discusses some additional technologies that are driving CR. Even though this is not a comprehensive list of driving technologies, it includes the most important ones.

2.4b.5.1 Geolocation

Geolocation is an important CR enabling technology due to the wide range of applications that may result from a radio being aware of its current location and possibly being aware of its planned path and destination.

The GPS is a satellite-based system that uses the time difference of arrival (TDoA) to geolocate a receiver. The resolution of GPS is approximately 100 m. GPS receivers typically include a one-pulse-per-second signal that is Kalman filtered as it arrives at each radio from each satellite, resulting in a high-resolution estimate of propagation delay from each satellite regardless of position. By compensating each pulse for the predicted propagation delay, the GPS receivers estimate time to approximately 340 nanoseconds (ns, or 10^{-9} seconds) of jitter [31].

In the absence of GPS signals, triangulation approaches may be used to geolocate a radio from cooperative or even non-cooperative emitters. Literature is available to discuss the classical approaches of TDoA, time of arrival (ToA), and, if the hardware supports it, angle of arrival (AOA). Multiple observations from multiple positions are required to create an accurate location estimate. The circular error probability (CEP) characterizes the estimate accuracy.

2.4b.5.2 Spectrum awareness/frequency occupancy

A radio that is aware of spectrum occupancy may exploit this information for its own purposes, such as utilization of open channels on a non-interference basis. Methods for measuring spectrum occupancy are discussed in commonly available literature.

A simple sensor resembles a spectrum analyzer. The differences are in quality and speed. The CR application must consider the quality of the sensor in setting parameters such as maximum time to vacate a channel upon use by an incumbent signal. It ingests a band of interest and processes it to detect the presence of signals above the noise floor. The threshold of energy at which occupancy is declared is a critical parameter. The detected energy is a function of the instantaneous power, instantaneous bandwidth, and duty cycle. An unpredictable duty cycle is expected. Spectrum occupancy is spatially variant, time variant, and subject to observational blockage (e.g., deep fading may yield a poor observation). Therefore, a distributed approach to spectrum sensing is recommended.

The primary problem associated with spectrum awareness is the hidden node problem. A lurking receiver (the best example is a television (TV) set) may be subjected to interference and may not be able to inform the CR that its receiver is experiencing interference. Regulators, spectrum owners, and developers of CR are working to find robust solutions to the hidden node problem. Again, a cooperative approach may help to mitigate some of the hidden node problems, but a cooperative approach will not necessarily eliminate the hidden node problem.

In addition to knowing the frequency and transmit activity properties of a radio transmitter, it may also be desirable for the radio to be able to recognize the waveform properties and determine the type of modulation, thereby allowing a radio to request entry into a local network. Many articles have been published on this topic, as well as a textbook by Azzouz and Nandi [32]. Once the modulation is recognized, then the CR can choose the proper waveform and protocol stack to request entry into the local network.

2.4b.5.3 Biometrics

A CR can learn the identity of its user(s), enabled by one or more biometric sensors. This knowledge, coupled with authentication goals, can prevent unauthorized users

Biometrics in Order of Effectiveness	Biometrics in Order of Social Acceptability
1. Palm scan 2. Hand geometry 3. Iris scan 4. Retina scan 5. Fingerprint 6. Voiceprint 7. Facial scan 8. Signature dynamics 9. Keyboard dynamics	1. Iris scan[a] 2. Keyboard dynamics 3. Signature dynamics 4. Voiceprint[b] 5. Facial scan[a] 6. Fingerprint[c] 7. Palm scan[c] 8. Hand geometry[c] 9. Retina scan[a]
	[a]Requires a camera scanner [b]Utilizes a copy of the voice input (low impact) [c]Requires a sensor in the push-to-talk (PTT) hardware

Figure 2.4b-3 Biometric sensors for CR authentication applications. Several biometric measures are low impact ones in terms of user resistance for authentication applications.

from using the CR. Most radios have sensors (e.g., microphones) that may be used in a biometric application. Voice print correlation is an extension to an SDR that is achievable today. Requirements for quality of voice capture and signal processing capacity are, of course, levied on the radio system. The source radio can authenticate the user and add the known identity to the data stream. At the destination end, decoded voice can be analyzed for the purposes of authentication.

Other biometric sensors can be used for CR authentication and access control applications. Traditional handsets may be modified to capture the necessary inputs for redundant biometric authentication. For example, cell phones recently have been equipped with digital cameras. This sensor, coupled with facial recognition software, may be used to authenticate a user. An iris scan or retina scan is also possible. Figure 2.4b-3 shows some of the potential sensors and their relative strengths and weakness in terms of reliability and acceptability [33].

2.4b.5.4 Time

Included in many contracts is the phrase "time is of the essence," testament to the criticality of prompt performance in most aspects of human interaction. Even a desktop computer has some idea about what time it is, what day it is, and even knows how to utilize this information in a useful manner (date and time stamping information). A radio that is ignorant of time has a serious handicap in terms of learning how to interact and behave. Therefore, it is important for the CR to know about time, dates, schedules, and deadlines.

Time-of-day information enables time division multiplexing on a coarsegrained basis, or even a fine-grained basis if the quality of the time is sufficiently accurate. Time-of-day information may gate policies in and out. Additionally, very fine knowledge of time may be used in geolocation applications.

GPS devices report time of day and provide a one-pulse-per-second signal. The one-pulse-per-second signal is transmitted from satellites, but does not arrive at every GPS receiver at the same time due to differences in path lengths. A properly designed receiver will assess the propagation delay from each satellite and compensate each of these delays so that the one-pulse-per-second output is synchronous at all receivers with only a 340 ns jitter. This level of accuracy is adequate for many applications, such as policy gating and change of cryptographic keys. Increased accuracy and lowered jitter may be accomplished through more sophisticated circuitry.

The local oscillator in a radio system may be used to keep track of time of day. The stability of these clocks is measured at approximately 10^{-6}. These clocks tend to drift over time, and in the course of a single day may accumulate up to 90 ns of error. Atomic clocks have much greater stability (10^{-11}), but have traditionally been large and power hungry. Chip-scale atomic clocks have been demonstrated and are expected to make precision timing practical. This will enable geolocation applications with lower CEPs.

2.4b.5.5 Spatial awareness or situational awareness

A very significant role for a CR may be viewed as a personal assistant. One of its key missions is facilitating communication over wireless links. The opposite mission is impeding communications when appropriate. As an example, most people do not want to be disturbed while in church, in an important meeting, or in a classroom. A CR could learn to classify its situation into "user interruptible" and "user non-interruptible." The radio accepting aural inputs can classify a long-running exchange in which only one person is speaking at a time as a meeting or classroom and autonomously put itself into vibration-only mode. If the radio senses its primary user

is speaking continuously or even 50 percent of the time, it may autonomously turn off the vibration mode.

2.4b.5.6 Software technology

Software technology is a key component for CR development. This section discusses key software technologies that are enabling CR. These topics include policy engines, artificial intelligence (AI) techniques, advanced signal processing, networking protocols, and the Joint Tactical Radio System (JTRS) Software Communications Architecture (SCA).

Policy engines

Radios are a regulated technology. A major intent of radio regulatory rules is to reduce or avoid interference among users. Currently, rules regarding transmission and reception are enumerated in spectrum policy as produced by various spectrum authorities (usually in high-level, natural language). Regulators insist that even a CR adhere to spectrum policies. To further complicate matters, a CR may be expected to operate within different geopolitical regions and under different regulatory authorities with different rules. Therefore, CRs must be able to dynamically update policy and select appropriate policy as a function of situation.

Spectrum policies relevant to a given radio may vary in several ways:

1. Policies may vary in time (e.g., time of day, date, and even regulations changing from time to time).
2. Policies may vary in space (e.g., radio and user traveling from one policy regulatory domain to another).
3. A spectrum owner/leaser may impose policies that are more stringent than those imposed by a regulatory authority.
4. The spectrum access privileges of the radio may change in response to a change in radio user.

As a result, the number of different policy sets that apply to various modes and environments grows in a combinatorial fashion. It is impractical to hard-code discrete policy sets into radios to cover every case of interest. The accreditation of each discrete policy set is a major challenge. SDRs, for example, would require the maintenance of downloadable copies of software implementations of each policy set for every radio platform of interest. This is a configuration management problem.

A scalable expression and enforcement of policy is required. The complexity of policy conformance accreditation for CRs and the desire for dynamic policy lead to the conclusion that CRs must be able to read and interpret policy. Therefore, a well-defined, accepted (meaning endorsed by an international standards body) language framework is needed to express policy. For example, if an established policy rule is constructed in the presence of other rules, the union of all policies is applicable. This enables hierarchal policies and policy by exception. As an example, suppose the emission level in band A is X dBm, except for a sub-band A′ for which the emission-level constraint is Y dBm if a Z KHz guard band is allowed around legacy signals. Even this simple structure is multi-dimensional. Layers of exceptions are complex. The policy engine must be able to constrain behavior according to the intent of the machine-readable policy. An inference capability is needed to interpret multiple rules simultaneously.

In the case of spectrum subleasing, policies must be delegated from the lessor to the lessee, and a machine-readable policy may be delegated. When a CR crosses a regulatory boundary, the appropriate policy must be enabled. Policies may also be used by the system in a control function.

The policy should use accepted standard tools and languages because the policy engine must be able to access automatic interpretation to achieve the goals of CR applications. Policies may be written by regulatory agencies or by third parties and approved by regulators, but in all cases policy is a legal or contractual operating requirement and provability in the policy interpretation engine is needed for certification.

Several suggestions for policy language have emerged. The Extensible Markup Language (XML) is not appropriate because it does not typically have inference capabilities in the interpretation engines. The Ontology Inference Layer (OIL), Web Ontology Language (OWL), and DARPA Agent Markup Language (DAML) have all been explored as possible policy languages. DARPA's XG program cites the OWL language as an appropriate language. Tool sets are available for building policy definitions and for machine interpretation of the definitions [34].

AI Techniques

The field of AI has received a great deal of attention for decades. In 1950, Alan Turing proposed the Turing test, regarding interacting with an entity and not being able to tell if it is human or machine. The AI techniques that work are plentiful, but most are not widely applicable to a wide range of problems. The powerful techniques may even require customization to work on a particular problem.

An agent is an entity that perceives and acts. A smart agent model is appropriate for CR. Figure 2.4b-4 explains four models of smart agents—simple reflex agents, model-based reflex agents, goal-based agents, and utility-based agents—defined as follows:

1. A *simple reflex agent* is a simple mapping from current sensor inputs to actuator settings. This is

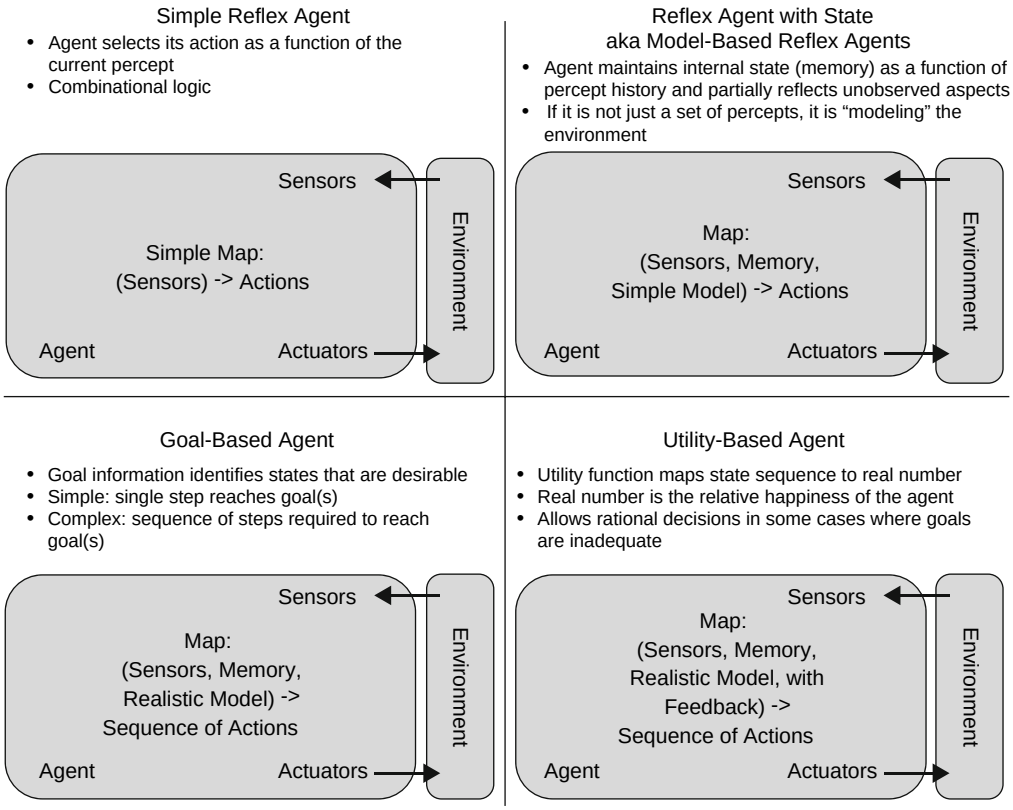


Figure 2.4b-4 Four smart agent models. Smart agents provide a framework that is consistent with the continuum of software maturity in software radios.

a stateless agent model that neither learns nor adapts to the environment.

2. A *model-based reflex agent* is still a simple mapping but now includes memory of past inputs. The actions are a function of the current sensor inputs and the recent past inputs, making it a finite memory agent model. There is still no learning. Adaptation is limited, but this is the minimum level of sophistication for an adaptable radio.
3. A *goal-based agent* adds to the memory of past inputs; it is a "realistic" model of the environment. Now a sequence of actions may be "tested" against a goal and an appropriate next action may be selected. The level of sophistication for the model of the environment is not well defined. These agents have increased capability of adapting because a prediction about the consequences of an action is available. There is no feedback, and learning is therefore limited. This is the minimal level of sophistication for a CR.
4. A *utility-based agent* maps the state sequence (memory state) to a "happiness" value and therefore includes feedback. The more sophisticated environment model may experiment with sequences of actions for selection of the best next action. This model of a CR has the ability to learn and adapt [35].

A smart agent model for CR is appropriate. The agent framework supports the continuum of radio maturity, and it allows the modular introduction of various AI techniques from fuzzy control to genetic algorithms (GAs). Agents may be tailored to the application's environment. In this sense, environment may be characterized in the following dimensions: fully observable versus partially observable, deterministic versus stochastic, episodic versus sequential, static versus dynamic, discrete versus continuous, and single agent versus multiagent.

The following is an incomplete list of AI techniques likely to find applicability to CR:

- State space models and searching
- Ontological engineering
- Neural networks in CR
- Fuzzy control in CR
- GAs in CR
- Game theory in CR
- Knowledge-based reasoning in CR

Signal processing

Digital signal processing (DSP) technology enables rapid advances in CRs. Intellectual property resources are widely available for signal processing functions. In GPP (general-purpose processor) or DSP resources, libraries of

routines realize functions in efficient assembly language. In FPGA (field programmable gate array) or ASIC resources, licensable cores of signal processing engines are available in very-high-speed integrated circuit (VHSIC) Hardware Design Language (VHDL) or Verilog. Signal processing routines are available for communication signal processing (modulation/demodulation, FEC, equalization, filtering, and others); audio signal processing (voice coding, voice generation, natural language processing); and sensor signal processing (video, seismic, chemical, biometric, and others).

Synthesizing signal processing functions together to form a system is a complex task. A process for algorithm development, test case generation, realization, verification, and validation eases the process of building a waveform or a cognitive system. Integrated tools for system development are available. Many of the tool sets include automatic generation of high-level language source code or hardware definition language code. A bit-level accurate simulation environment is used to develop the system algorithms and to generate test cases for post-integration verification and validation. This environment may be used for CRs to synthesize communications or multimission waveforms that enable a CR to achieve specific goals.

Networking protocols

Cooperative groups (a multiagent model) have the potential to increase capabilities in a variety of ways. For example, a lone CR is limited in its ability to access spectrum, but a pair of CRs can sense the spectrum, identify an unused band, rendezvous there, and communicate. A network of CRs enables other significant increases in capabilities. Software for Mobile Ad hoc Networking (MANET), although maturing slowly, is a key enabling technology.

The medium access control (MAC) layer is critical in CR networks. If the CR is employing advanced spectrum access techniques, a robust MAC that mitigates the hidden node problem is needed. In a "static spectrum access" environment, a more traditional MAC is possible. A carrier sense collision detection (802.11 MAC) mode is not possible because a radio cannot receive in the exact same band in which it is transmitting, so a carrier sense collision avoidance approach is frequently used. Request-to-send (RTS) and clear-to-send (CTS) messaging are popular wireless MACs amenable to MANETs. Other approaches include TDMA or CDMA MACs.

The architecture for routing packets is important for performance in MANETs. The approaches are generally divided into proactive and reactive algorithms. In a proactive-routing environment, routing data, often in the form of a routing table, are maintained so that a node has a good idea of where to send a packet to advance it toward its final destination, and a node may know with great confidence how to route a packet even if one is not ready to go. Maintaining this knowledge across a MANET requires resources. If the connection links are very dynamic or the mobility of the nodes causes rapid handoff from one "network" to another, then the overhead to maintain the routing state may be high. In contrast, reactive-routing approaches broadcast a short search packet that locates one or more routes to the destination and returns that path to the source node. Then the information packet is sent to the destination on that discovered route. This causes overhead in the form of search packets. Proactive and reactive routing both have pros and cons associated with their performance measures, such as reliability, latency, overhead required, and so on. A hybrid approach is often best to provide scalability with offered network load.

An interesting application of CR is the ability to learn how to network with other CRs and adapt behavior to achieve some QoS goal such as data rate below some BER bound, bounded latency, limited jitter, and so forth. Various cognitive-level control algorithms may be employed to achieve these results. As an example, a fixed-length control word may be used to parameterize a communications waveform with frequency, FEC, modulation, and other measures. The deployment of a parameterized waveform may be controlled and adapted by using a generic algorithm and various QoS measures to retain or discard a generated waveform.

Software communications architecture

The primary motivations for SDR technology are lower life cycle costs and increased interoperability. The basic hardware for SDR is more expensive than for a single-point radio system, but a single piece of hardware performs as many radios. The single piece of hardware requires only one logistics tail for service, training, replacement parts, and so on. One of the driving costs in SDR development is that of software development. The JTRS acquisitions are controlling these costs by ensuring software reuse. The approach for reuse is based on the Software Communications Architecture (SCA), which is a set of standards that describes the software environment. It is currently in release 2.2.1. Software written to be SCA compliant is more easily ported from one JTRS radio to another. The waveforms are maintained in a JTRS Joint Program Office (JPO) library.

A CR can be implemented under the SCA standards. Applications that raise the radio to the level of a CR can be integrated in a standard way. It is expected that DARPA's XG program will provide a CR application for policy-driven, dynamic spectrum access on a non-interference basis that executes on JTRS radios. XG is the front-runner in the race to provide the first military CR.

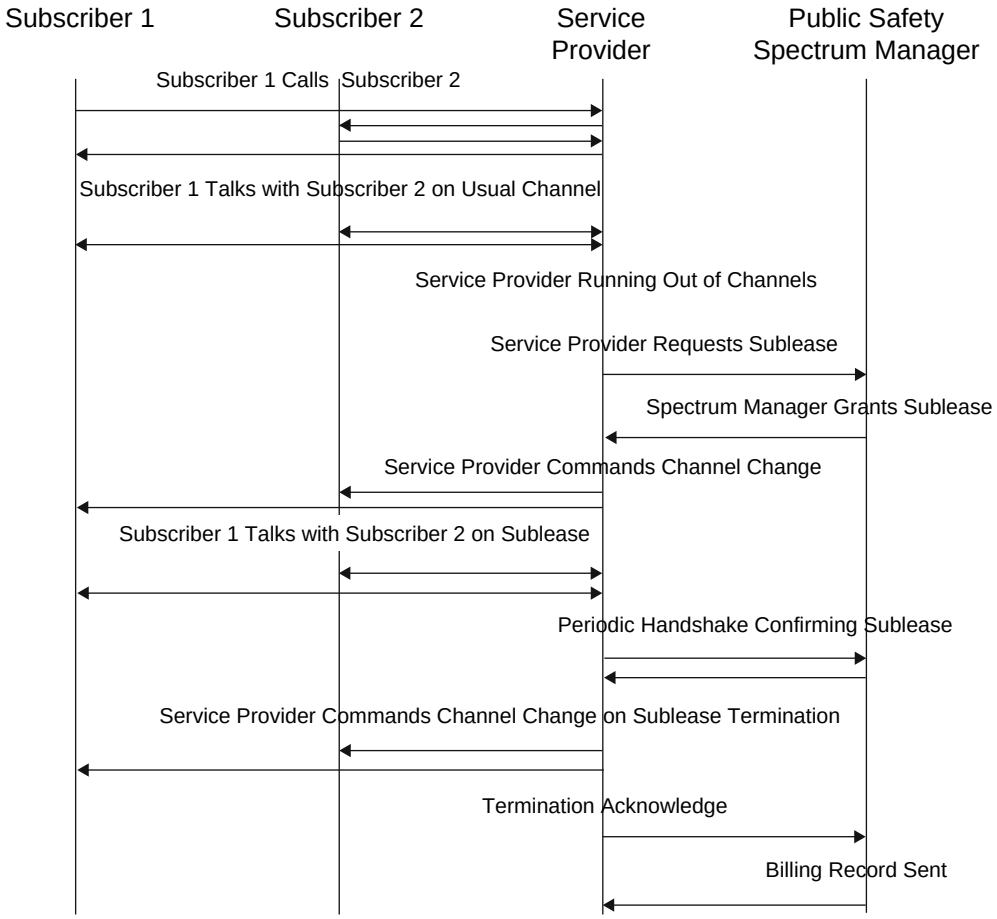


Figure 2.4b-5 Spectrum subleasing sequence. One motivation for CR is the potential income stream derived from subleasing idle spectrum on a non-interference basis.

2.4b.5.7 Spectrum awareness and potential for sublease or borrow

The Spectrum Policy Task Force (SPTF) recommends that license holders in exclusive management policy bands be allowed to sublease their spectrum. Figure 2.4b-5 shows a sequence diagram for spectrum subleasing from a public safety spectrum owner. During the initial contact between the service provider and the public safety spectrum manager, authentication is required. This ensures that spectrum use will be accomplished according to acceptable behaviors and that the bill will be paid [36].

For a subleasing capability to exist in a public safety band a shut-down-on-command function must be supported with a bounded response time. There are three approaches to this: continuous spectrum granting beacon, time-based granting of spectrum, and RTS–CTS–inhibit send. Figure 2.4b-5 shows the time-based granting of spectrum, described as a periodic handshake confirming sublease.

Even though public safety organizations may not sublease spectrum, other organizations may choose to do so. Subleasing has the benefit of producing an income stream from only managing the resource. Given proper behavior by lessees and lessors, the system may become popular and open up the spectrum for greater utilization.

2.4b.6 Funding and research in CRs

DARPA is funding a number of cognitive science applications, including: the XG Program, Adaptive Cognition-Enhanced Radio Teams (ACERT), Disruption Tolerant Networking (DTN), Architectures for Cognitive Information Processing (ACIP), Real World Reasoning (REAL), and DAML. DARPA research dollars under contract are easily in the tens of millions. Good results have been achieved in many of these efforts.

The National Science Foundation (NSF) is also funding cognitive science applications including grants to the Virginia Polytechnic Institute and State University (Virginia Tech or VT). Additionally, NSF has sponsored

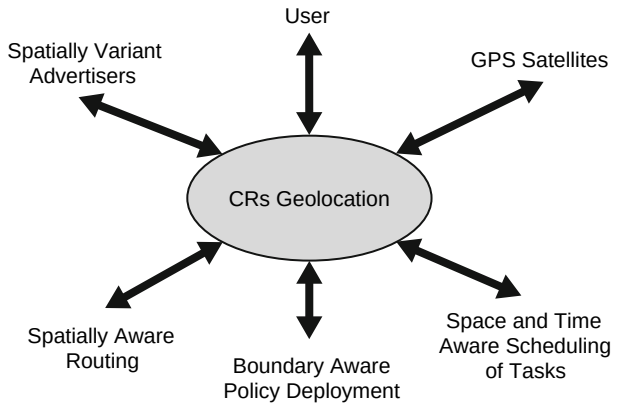


Figure 2.4b-6 CR use case for geolocation. Several interactions between the CR and the world are enabled or required by geolocation information.

Information Theory and Computer Science Interface workshops that communicate CR research results.

The SDR Forum has a CR Working Group that is investigating various CR technologies such as spectrum access techniques, and a Cognitive Applications Special Interest Group that is working with regulatory bodies, spectrum owners, and users to communicate the potential of CR technologies. Numerous organizations participate at SDR Forum meetings. It is expected that the SDR Forum will solicit inputs from industry through a Request for Information (RFI) process in the near future.

Both the Federal Communications Commission (FCC) and the National Telecommunications and Information Administration (NTIA) have interest in CR. The FCC has solicited various comments on rule changes and an SPTF report. NTIA has been involved in the discussions as they relate to government use of spectrum.

2.4b.6.1 Cognitive geolocation applications

If a CR knows where it is in the world, myriad applications become possible. The following is a short set of examples. Figure 2.4b-6 shows a use case level context diagram for a CR with geolocation knowledge.

Establishing the location of a CR enables many new functions. A cognitive engine for learning and adapting can utilize some of the new functions. There are multiple methods for determining location. For example, GPS receiver technology is small and inexpensive. Given an appropriate API, numerous applications including network localization (discussed next) and boundary awareness are available to the cognitive engine in the CR.

Network localization is a term to describe position aware networking enhancement. For example, if a radio knows it is in a car, that it is morning rush hour, and that the path being taken is the same as that of the last 200 days of commuting, it can predict being in the vicinity of the owner's office wireless local area network (LAN) within a certain time period. The radio may wait until it is in the office to download e-mail and upload the pictures of the accident taken a few minutes ago. This is an example of the highest level of management of radio functions (assuming the radio is used for e-mail, photos, etc.).

Spatial awareness may be used for energy savings. In a multiple short hop-routing algorithm with power management just closing the wireless link with multiple short hops is usually more energy efficient than one long hop. Knowing the position of each node in a potential route allows the CR to take energy consumption into consideration when routing a packet.

Spatially variant searching is a powerful concept for increasing a user's operating efficiency. If it is time for supper, the CR may begin a search for the types of restaurants the user frequents and sort them by distance and popularity. Other spatially variant searches are possible.

A radio aware of boundaries may be able to invoke policy as a function of geopolitical region. When passing from one regulatory jurisdiction to another, the rules change. The radio can adopt a conservative operation mode near the boundaries and change when they are crossed. The radio must have the ability to distinguish one region from another. This may require a database (which must be kept up-to-date) or some network connectivity with a boundary server. Figure 2.4b-7 shows a simplified sequence diagram in which a CR accesses spectrum as a function of a spatially variant regulatory policy.

A sequence of position estimates may be used to estimate velocity. Take a scenario in which teenagers' cell phones would tattle to their parents when their speed exceeds a certain speed. For example, suppose tattle mode is set. The radio is moving at 45 mph at 7:30 a.m. The radio calls the parent and asks a couple of questions such as: "Is 45 mph okay at this time? Is time relevant?" The questions will be in the same order each time and the parent won't have to wait for the whole question, just the velocity being reported. Then the parent keys in a response. After a few reports, the radio will develop a threshold as a function of time. For example, during the time the radio is heading for school, 45 mph is okay. During the lunch break (assuming a closed campus), 15 mph might be the threshold. An initial profile may be programmed, or the profile may be learned through tattling and feedback to the reports. Vehicle position and velocity might also be useful after curfew.

The CR application uses special hardware or customized waveforms that return geolocation information. This information is used to access databases of policies or resources to make better decisions. Dynamic exchange

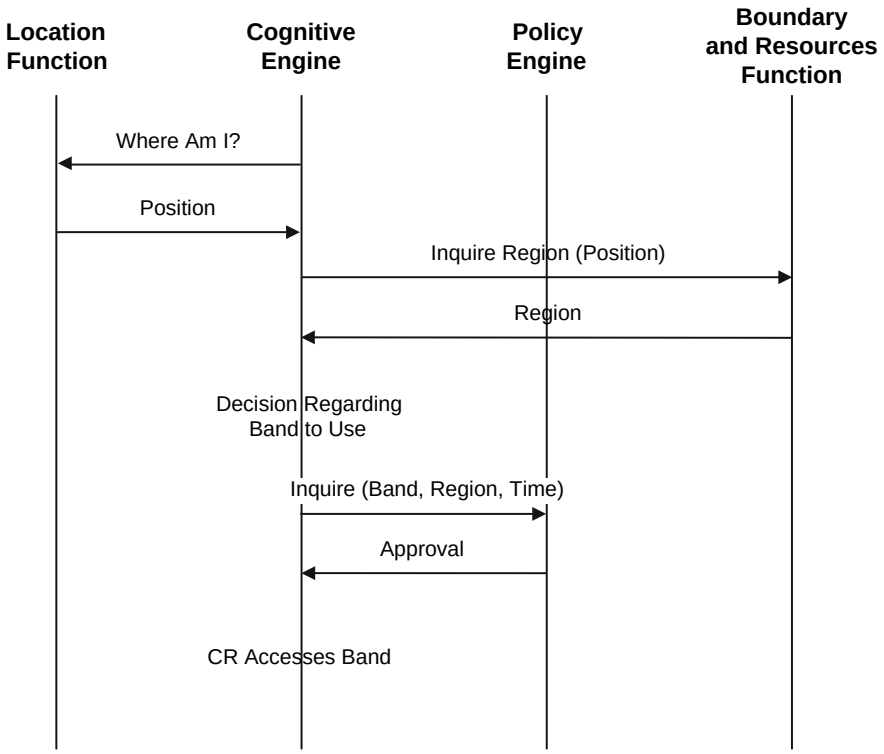


Figure 2.4b-7 Spectrum access sequence diagram. A policy engine uses regional inputs to select spatially variant policies to approve or disapprove a requested spectrum access.

of information may be used for other networking actions. The set of CR applications that are enabled by geolocation capability is large and has many attractive benefits.

2.4b.6.2 Dynamic spectrum access and spectrum awareness

One of the most common capabilities of CRs is the ability to intelligently utilize available spectrum based on awareness of actual activity. Current conservative spectrum management methods (static spectrum assignments) are limited because they reduce spatial reuse, preclude opportunistic utilization, and delay wireless communication network deployment. Without the need to statically allocate spectrum for each use, however, networks can be deployed more rapidly. A CR with spectrum sensing capability and cooperative opportunistic frequency selection is an enabling technology for faster deployment and increased spatial reuse.

Spectrum access is primarily limited by regulatory constraints. Recent measurements show that spectrum occupancy is low when examined as a function of frequency, time, and space [36]. CRs may sense the local spectrum utilization either through a dedicated sensor or by using a configured SDR receiver channel. Uses of this information may create increased spectrum access opportunities.

One of the primary considerations for such a cognitive application is noninterference with other spectral uses. Figure 2.4b-8 shows local spectrum awareness and utilization. If the regulatory body is allowing CRs to utilize the unoccupied "white space," increased spectral access can be achieved. The CR can examine the signals and may extract detailed information regarding use. By estimating the other uses and monitoring for interference, two CRs may rendezvous at an unoccupied "channel" and communicate.

Sophisticated waveforms that have the ability to periodically stop transmitting and listen for legacy users are called for in this application, as well as waveforms that can adapt their spectral shape over time. Dynamic selection of channels to utilize or vacate is important. Simulations of these cooperating CRs already exist, and additional field demonstrations are expected in the near future. Another advantageous waveform characteristic is discontinuous spectrum occupancy. This allows a wideband communication system to aggregate available spectrum between other existing signals. Careful analysis is needed to ensure that sufficient guard bands are utilized.

Figure 2.4b-9 shows five suggested alternatives for utilizing spectrum in and around legacy signals. The characteristics of the legacy signals may be provided to

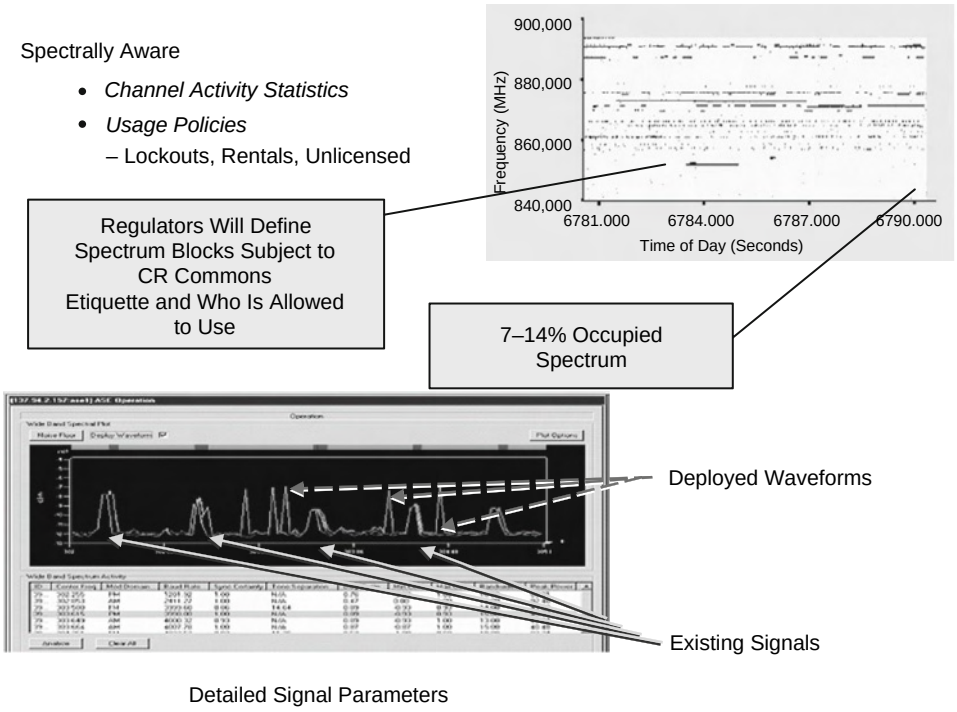


Figure 2.4b-8 Spectrum awareness. A CR, or a set of CRs, may be aware of the spectrum and may exploit unoccupied spectrum for its own purposes.

the cooperating CRs by federated sensors. An alternative method for characterizing the legacy signals is time division sharing of a channel as a sensor and providing a look-through capability by duty-cycling transmit and monitor functions. The five methods shown in Figure 2.4b-9 avoid the legacy signal in various ways.

The first method assumes a fixed bandwidth waveform: the center frequency may be adapted. The second method assumes a variable bandwidth signal, such as a direct-sequence spread-spectrum waveform, where the chip rate is adapted and the center frequency is adapted. The third method uses a water-filling method to populate a subset of the carriers in an OFDM waveform. These three methods impact legacy signals very little if appropriate guard bands are observed. The fourth method is a direct-sequence spread-spectrum waveform that underlies the legacy signals. The interference from this underlay must be very small so that legacy systems do not experience noise; thus the processing gain of the spread spectrum underlay must be very high. The last method shown also avoids the legacy signals by frequency hopping into only unoccupied channels.

The spread-spectrum method deserves some elaboration at this point. Legacy receivers will perceive the spread-spectrum signal as an increase in the noise floor, which may result in a decrease in the link margin of the legacy signal. The spread-spectrum receiver will perceive the legacy waveform(s) as a narrowband interference. The de-spreading of the desired signal will cause the narrowband signals to spread. This spreading will cause the narrowband signals to appear at a signal level reduced in power by the spreading gain, and thus appear as noise to the CR, resulting in reduced link margin. The ability of each of these communication systems to tolerate this reduced link margin is link specific and therefore a subject of great concern to the legacy system operators.

An OFDM waveform, the third method in Figure 2.4b-9, has several benefits including flat fading subchannels, elimination of equalizer due to long symbol time and guard intervals, the ability to occupy a variety of bandwidths that "fit to the available opportunity," and the ability to null subcarriers to mitigate interference. Variable bit loading enables pre-nulling and dynamic nulling. Table 2.4b-5 compares several methods for dynamic bit loading an OFDM waveform. Not loading subcarriers occupied by legacy signals with guard bands around them minimizes interference between CR and non-CR systems [37–39].

Because spectrum utilization is a spatially and temporally variant phenomenon, it requires repeated monitoring and needs cooperative, distributed coordination. The familiar hidden node and exposed node problems have to be considered. Figure 2.4b-10 shows a context diagram in which external sensor reports are made available to the CR and may be considered when selecting unoccupied bands.

Figure 2.4b-11 shows a sequence diagram in which a set of CRs is exchanging sensor reports and is learning

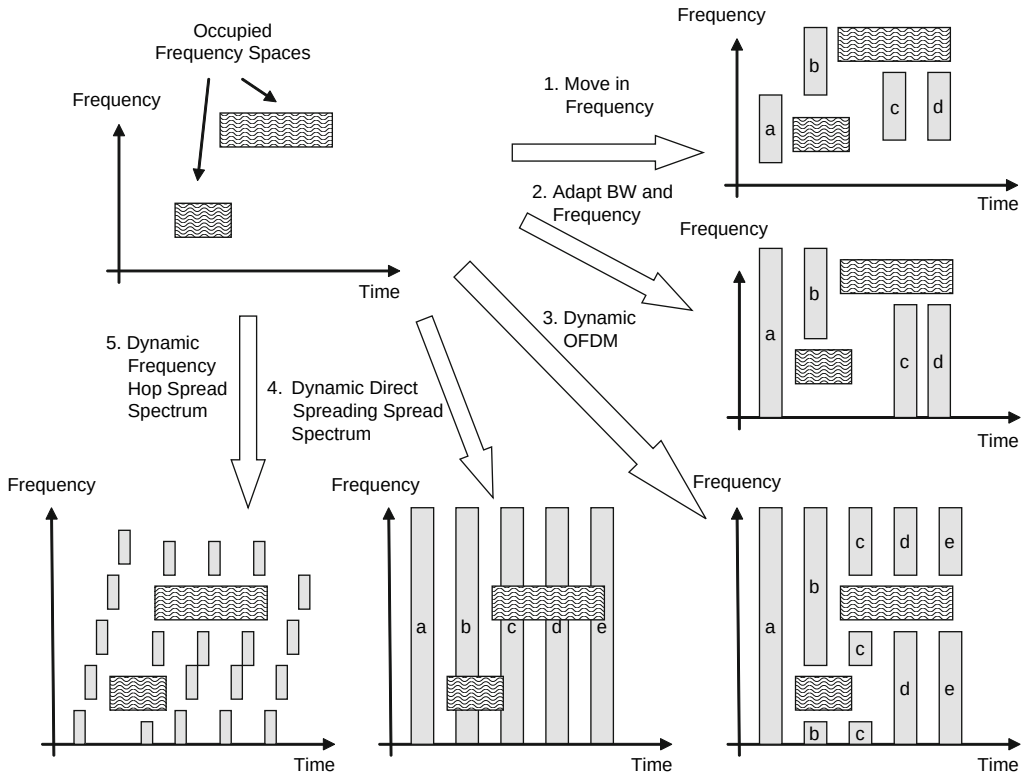


Figure 2.4b-9 Non-interference methods for dynamic spectrum access. Different strategies for deploying non-interfering waveforms have been proposed.

about local spectrum occupancy. At some time, a pair of CRs wishes to communicate and rendezvous at a band for that purpose. When a legacy signal is detected, the pair of CRs must vacate that band and relocate to another.

The sensor technology utilized for spectrum awareness should be of high quality to mitigate the hidden node problem. For example, if a CR wants to use a TV band, its sensor should be significantly more sensitive than a TV set so that if it detects a TV signal, it will not interfere with local TV set reception, and if it does not detect a signal, there is a high probability that no TV set is near enough to demodulate a TV signal on that channel.

Dynamic spectrum access as a result of learning about the spectrum occupancy is a strong candidate for a CR application. Numerous discussions in regulatory organizations have involved whether to allow this behavior. If implemented correctly, it is a win–win situation in which more spectrum is utilized and very little additional interference is suffered.

Table 2.4b-5 Comparison of variable bit loading algorithms.

Method	Characteristic	Complexity
Water filling	Original approach, optimal, frequently used for comparison	$O(N^2)$
Hughes-Hartogs [37]	Optimal, loads bits serially based on subcarrier energy level, slow to converge, repeated sorts	$O(SN^2)$
Chow [38]	Suboptimal, rounds to integer rates using signal-to-noise gap approximations, some sorting required	$O(N \log N + N \log S)$
Lagrange (unconstrained) Krongold [39]	Optimal, computationally efficient, efficient table lookup, Lagrange multiplier with bisection search, integer bit loading, power allocation, fast convergence	$O(N \log N)$, revised $O(N)$

Note: *N:* number of subcarriers; *S:* number of bits per subcarrier.

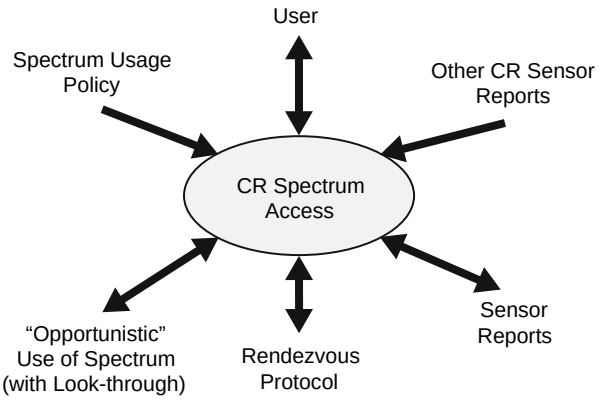


Figure 2.4b-10 Spectrum access context diagram. A spatially diverse sensing protocol is required to mitigate such problems as hidden node or deeply fading RF channels for CR access to spectrum on a non-interference basis.

2.4b.6.3 The rendezvous problem

The difficulty for the CR network is the radios locating each other and getting the network started. Before any transmission can occur, all radios must survey the spectrum to determine where the available "holes" (frequency reuse opportunities) are located. However, each receiver–sensor will perceive the spectrum slightly differently because each will see different objects shadowing different transmitters, and will either see each transmitter with a different signal strength or will not see some transmitters that other nodes are able to see. So while node A may see an open frequency, node B may consider that frequency to be in use by a distant transmitter node or network L. To get the network started, the CRs must agree on a protocol to find each other.

There are several possible methods by which to do this. These methods depend strongly on whether there is an infrastructure in place to help start up a CR network, or whether no infrastructure can be assumed.

Infrastructure-aided rendezvous

If we assume that there is an infrastructure component, we must also assume it behaves just like the CR and does not interfere with legacy systems. We assume, however, that it periodically transmits a beacon signal, and we assume that this beacon includes reference time, next frequency-hop(s), and a description of frequencies in use within the local region. Furthermore, we assume that the infrastructure beacon is followed by an interval in which CRs request an available time-frequency slot and associate a net name to that slot, as well as to their location and transmit power, followed by a response from the infrastructure recommending which frequency to use and when to check back. Subsequent requests by other net members can be directed to the proper time-frequency slot, and the geographic distribution of net members can be tracked, allowing the infrastructure to assess interference profiles.

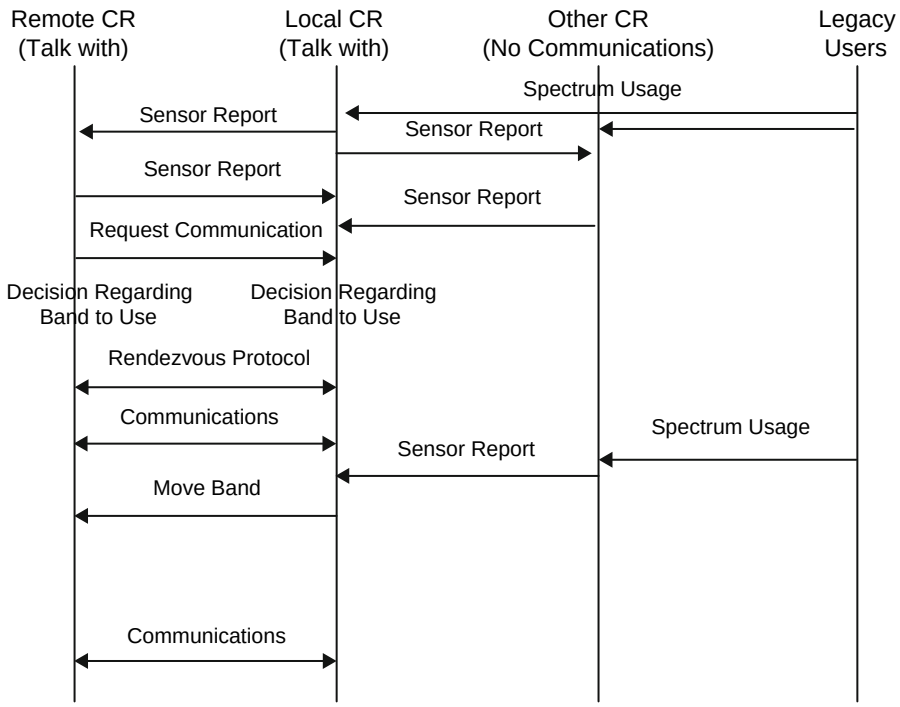


Figure 2.4b-11 Dynamic spectrum access sequence diagram (assumes a control channel). A pair of CR communicating on an unoccupied channel must vacate when interference potential is detected.

Unaided rendezvous

Defense systems are rarely able to assume support infrastructure. Similarly, early deployments of commercial CR equipment will not be able to assume infrastructure. Consequently, it is important to have an unaided method for rendezvous. Several methods exist for systems to find each other.

The problem is somewhat like two men searching for each other when they are a mile apart on a moonless dark night out in the desert. Each has a flashlight the other may see, but only when it is pointed in the right direction. They can finally signal each other, but only when each has noticed the other's flashlight, so each must look in the proper direction at the proper time.

In the case of CRs trying to find each other, one must transmit and the other must receive in the proper frequency "hole" at the proper time. Several methods are feasible. All involve one node transmitting "probe" signals (a uniquely distinguishing waveform that can be readily correlated) until receiving a response from a net member. In all cases, the problem is that the frequencies the transmitter sees as being "usable holes" are different from those the receiver sees. Thus, neither can be stationary and just sit on one frequency hoping the other will find it:

C*ase 1*: Node A transmits probes in randomly selected frequency holes, while node B listens to random frequencies considered to be unoccupied. Preferably, node B listens about three times as long as a probe pulse. The search time can be dramatically reduced if node B is able to listen to many simultaneous frequencies (as it might with an OFDM receiver). Node B responds with a probe response when node A is detected.

C*ase 2:* Node A selects the five largest unoccupied frequency blocks, and rotates probe pulses in each block. Similarly, node B scans all of the unoccupied frequency blocks that it perceives, prioritized by the size of the unoccupied frequency block, and with a longer dwell time.

Both of these cases are similar, but minor differences in first acquisition and robustness may be significant.

After node B hears a probe and responds with a probe acknowledge, the nodes will exchange each node's perception of locally available frequencies. Each node will "logical OR" active frequencies, leaving a frequency list that neither node believes is in use. Then node A will propose which frequency block to use for traffic for the next time slot, at what data rate, and in what waveform.

From this point forward, the nodes will have sufficient connectivity to track changes in spectral activity and to stay synchronized for which frequencies to use next.

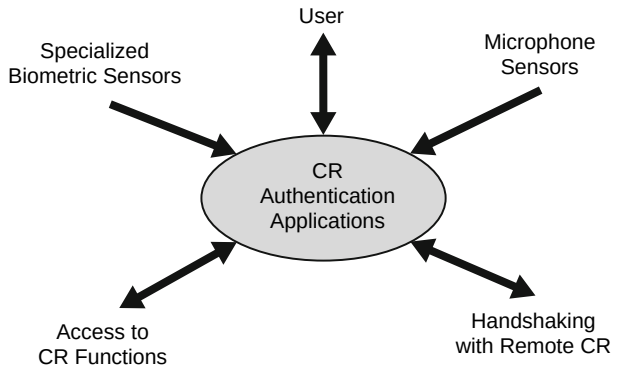


Figure 2.4b-12 Authentication context diagram. A CR application that learns the identity of its user(s) and has an Access List Security Protocol can minimize fraud and maximize secure communications.

2.4b.6.4 CR Authentication applications

A CR can learn the identity of its user(s). Authentication applications can prevent unauthorized people from using the CR or the network functions available to the CR. This enhanced security may be exploited by the military for classified communications or by commercial vendors for fraud prevention.

Because many radios are usually used for voice communications, a microphone often exists in the system. The captured signal is encoded with a VoCoder (voice coder) and transmitted. The source radio can authenticate the user (from a copy of the data) and add the known identity to the data stream. At the destination end, decoded voice can be analyzed for the purposes of authentication, and the result may be correlated with the sent identity.

Other sensors may be added to a CR for the purposes of user authentication. A fingerprint scanner in a push-to-talk (PTT) button is not intrusive. Automatic fingerprint correlation software techniques are available and scalable in terms of reliability versus processing load required. Additionally, cell phones have been equipped with digital cameras. This fingerprint sensor coupled with facial recognition software may be used to authenticate a user. Again, the reliability is scalable with processor demands.

Figure 2.4b-12 shows a context diagram for a CR's authentication application. The detailed learning associated with adding a user to an access list through a "third-party" introduction is not shown. The certificate authority could be the third party, or another authorized user could add a new user with some set of authority. The CR will learn and adapt to the new set of users and to the changing biometric measures of a user. For example, if a user gets a cold, his voice may change, but the correlation between the "new voice" and the fingerprint

scanner is still strong, and the CR may choose to temporarily update the voice print template.

2.4b.7 Timeline for CRs

CR development will be a spiral effort. The when, where, how, and who of this development are discussed below:

When: Currently, several CR initiatives are under way. Progress is evident every day, and more and more sophisticated demonstrations are imminent. Some of these demonstrations will include better policy conformance characteristics.

Where: The FCC and NTIA are currently discussing a test band for CR. They are suggesting a 10 MHz experimental chunk of spectrum to allow developers to experiment. This promising development must be exploited. As policy is written for this band, it can be deployed and policy violations may be assessed.

Who: Vendors, regulators, service providers, and users are highly interested in CR systems. A great deal of discussion on exactly what that means has already taken place and continues. Academic researchers using COTS SDR demonstrations and government-sponsored demonstrations using custom-developed SDRs will reduce CR to practice, for at least the first systems. Where the technology will progress is difficult to predict. As an example, the following organizations are working with CR: General Dynamics; Shared Spectrum; Raytheon; Lockheed Martin; Bolt, Beranek, & Newman; Rockwell Collins; Harris; Virginia Tech; and Northeastern University (and the list has doubled during the preparation of this book).

How: The easiest experiments utilize COTS SDR hardware and execute cognitive applications as demonstrations. Custom-developed hardware is more expensive, but is better tailored to show the benefits of CR. Spectrum awareness using a spectrum sensor has the best information to exploit unoccupied spectrum. This custom capability also has computational resources sized to execute a policy constraint engine.

2.4b.7.1 Decisions, directions, and standards

Numerous organizations and standards bodies are working in the area of CR. The SDR Forum, the Institute of Electrical and Electronics Engineers (IEEE), the Federal Communications Commission (FCC), the National Telecommunications and Information Administration (NTIA), and the International Telecommunication Union (ITU) all have interests in this area.

2.4b.7.2 Manufacture of new products

Many products have new sensors and actuators in them. Cellular telephone handsets are a high-volume, highly competitive product area, from which innovation is driven by these characteristics. In fact, a large fraction of new cell phones have an integrated digital camera. These are manually operated today, but CR applications may take advantage of this sensor for more "cognitive" operation modes.

Chemical sensors have been integrated to some cell phone models. The purpose of the chemical sensors is to report on important "problems" such as "bad breath" and "blood alcohol level." This is a manually operated device, but future applications may be more autonomous.

Among the new applications in cellular telephones is a Bluetooth-like waveform to introduce single people. "Flirting radios" may subsequently need to have AI technology added to filter out the "undesirable" introductions.

Much to the dismay of teenagers and employees everywhere, phone tracking applications are now available. Although these capabilities are used "manually" now, learning algorithms can be applied to the interface to create a new "filtered" report stream. There is serious interest in tracking other sets of people, such as first responders, delivery people, service people, or doctors.

2.4b.8 Summary and conclusions

Radio evolution has taken the path toward more digital realizations and more software capabilities. The original introduction of software made possible software capable radios that communicate and process signals digitally. In the pursuit of flexibility, software programmable radios, and the even more flexible SDR, have become the standard in the military arena and are starting to gain favor in the commercial world, as explained in Section 2.4b.2. We are now seeing the emergence of aware radios, adaptive radios, and finally CRs, and we have traced the continuum among these various degrees of capability, as well as providing a few examples of each. Sections 2.4b.3 and 2.4b.4 explored the properties and capabilities of each of these classes of radios.

Section 2.4b.5 outlined the enabling technologies for CRs. Numerous technologies have matured to the point where CR applications are possible and even attractive. The ability to geolocate a system, sense the spectrum, know the time and date, sense biometric characteristics of people, access new software capabilities, and determine new regulatory environments are all working together to enable CR.

Geolocation through the use of GPS or other methods is now available. This enabling technology allows a radio

to make spatially variant decisions, which may be applied to the selection of policy or networking functions.

Sensing of the local RF environment is available. This information may be used to mitigate a deeply fading channel or may be used to access locally unoccupied spectrum. Non-interference is particularly important, and protocols for sensing, deciding, and accessing spectrum are being designed, developed, and demonstrated today.

Increased robustness in biometric sensor technology provides a whole new dimension to CR applications. The most likely initial use of this technology is in user authentication applications, such as the purchasing of services and products.

Knowledge of time has been available in many forms, but integration into a broader range of full-function capabilities will enable all new applications. Stable time knowledge enables a CR to plan and execute with more precision. Using this capability for non-infrastructure-based geolocation, dynamic spectrum access, or AI planning is envisioned for near-term CR functions.

A smart agent model of CRs is attractive. An agent is an entity that perceives and acts on behalf of another. This is where CRs are going. Smart agent models of the world enable a radio to provide services for its user. Improved performance or new capabilities may be provided. As the CR's smart agent model of the world becomes more sophisticated and realistic, situational awareness will increase. As the models improve, the ability of the CR to act effectively over a broader range of user services will improve.

Maybe the most important software technology is a policy engine that enables a CR to know how to behave where it is right now, given the priorities of current circumstances. AI applications at a very high level, networking services in the middle levels, and signal processing primitives at a very low level are all available for a CR developer to utilize in creating new capabilities. Finally, middleware technology enables greater software reuse, which makes CR development economical.

Modern regulatory philosophy is starting to allow CR to deploy new services and capabilities. As the trend continues, there are economic motivations for deploying CR systems.

Section 2.4b.6 covered research in CR technologies, and presented three significant classes of CR applications. Geolocation-enabled applications and authentication applications were discussed in some detail. The most promising CR application is dynamic spectrum access. Suggestions for using OFDM waveforms along with dynamic bit loading are included in this chapter. Solutions to the rendezvous problem are suggested, and the hidden node problem is described.

Section 2.4b.7 covered the timeline in which these technologies will roll out and be integrated into radio equipment and products. Many of the technologies required to provide some of the useful and economically important CR functions already exist, so some of these features should begin to appear within the timeline of the next development cycle.

The bottom line is that the enabling technology for CR applications is available. There is interest in integrating the technologies to build cognitive applications. Finally, the emergence of CRs and their cognitive applications is imminent.

References

[1] Federal Communications Commission, Notice of Proposed Rule Making, August 12, 2000.

[2] http://www.fas.org/man/dod-101/sys/land/sincgars.htm

[3] http://www.fas.Org/man/dod-101/sys/land/.htm

[4] http://www.fas.org/spp/military/program/com/an-wsc-3.htm

[5] http://www.fas.org/man/dod-101/sys/ac/equip/an-arc-164.htm

[6] http://www.fas.org/man/dod-101/sys/ac/equip/an-arc-220.htm

[7] http://www.harris.com

[8] http://www.columbiaelectronics.com/motorola_LST_5b___LST_bc.htm

[9] http://army-technology.com

[10] http://www2.thalescomminc.com

[11] http://www.fas.org/man/dod-101/sys/ac/equip/csel.htm

[12] www.jcs.mil/j6/cceb/jtidsmidswgnotebookjune2005.pdf

[13] http://www.fas.org/man/dod-101/sys/ac/equip/an-arc-210.htm

[14] http://www.afcea.org/signal

[15] http://www.fas.org/man/dod-101/sys/land/eplrs.htm

[16] http://www.nj7p.org/history/portable/html

[17] http://raytheon.com/products/jtt_cibs/

[18] http://www.acd.itt.com/case4.htm

[19] http://jtrs.army.mil/

[20] http://jtrs.army.mil/

[21] http://www2.thalescomminc.com

[22] www.comsoc.org/tech_focus/pdfs/rf/06.pdf

[23] http://www.gdc4s.com/

[24] http://jtrs.army.mil/

[25] http://www.gnu.org/software/gnuradio

[26] http://vanu.com/technology/softwareradio.html

[27] J. Polson, "Cognitive Radio Applications in Software Defined Radio," in *Software Defined Radio Forum Technical Conference and Product Exposition*, Phoenix, AZ, November 15–18, 2004.

[28] H. Chen, S. Kumar and C.-C. Jay Kuo, "QoS-Aware Radio Resource Management Scheme for CDMA Cellular Networks Based on Dynamic Interference Guard Margin (IGM)," *Computer Networks*, Vol. 46, 2004, pp. 867–879.

[29] B. Bougard, S. Pollin, G. Lenoir, L. Van der Perre, F. Catthoor and W. Dehaene, "Energy-Aware Radio Link Control for OFDM-Based WLAN." Available at http://www.homes.esat.kuleuven.be/~bbougard/Papers/sips04-l.pdf.

[30] http://www.darpa.mil

[31] P.H. Dana, *The Geographer's Craft Project*, Department of Geography, University of Colorado, Boulder, 1996.

[32] E. Azzouz and A. Nandi, *Automatic Modulation Recognition of Communication Signals*, Springer, New York, 1996.

[33] S. Harris, *CISSP All-in-One Exam Guide*, 2nd Edition, McGraw-Hill, New York, 2002.

[34] http://www.bbn.com

[35] S.J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd Edition, Pearson Education, London, 2003.

[36] Federal Communications Commission. *Spectrum Policy Task Force Report*. ET Docket No. 02–135, November 2002.

[37] Hughes-Hartogs, "Ensemble Modem Structure for Imperfect Transmission Media," United States Patent 4,679,227, July 7, 1987.

[38] P. Chow, J. Cioffi and J. Bingham, "A Practical Discrete Multitone Transceiver Loading Algorithm for Data Transmission Over Spectrally Shaped Channels," *Transactions on Communications*, Vol. 43, February /March/April (2/3/4), 1995, 773–775.

[39] B. Krongold, K. Ramchandran and D. Jones, *Computationally Efficient Optimal Power Allocation Algorithms for Multicarrier Communications Systems*, University of Illinois, Urbana–Champaign.

This page intentionally left blank

Chapter 2.5

2.5

Introduction to RF and microwave radiation

Ronald Kitchen

This chapter gives an outline of the essential aspects of transmission and the nature of electromagnetic waves particularly for those from other disciplines who may otherwise find the electronic content unfamiliar. It includes explanations of the terms relevant to this form of non-ionising radiation.

Radio frequency (RF) radiation

The previous book on this subject was entitled *RF Radiation Safety Handbook*, the term 'RF' covering all frequencies used for communications, radar, satellites, etc., up to the nominal ceiling of 300 GHz. However, it was suggested that some people regard radio frequency (RF) as applying only to the lower part of this spectrum. Consequently the word 'microwave' has been added in this revision, although it is redundant in the context of the book. It would be tedious to use both terms throughout the book so 'RF' is used to include 'microwaves' here as is understood by radio engineers. The term microwave is only specifically used when the topic involves something to which the term normally attaches, e.g. microwave oven, microwave antenna, etc.

The subject of RF radiation is still regarded as mysterious and something of a black art. This is no doubt due to the fact that it cannot be seen or touched. There was also an element of magic in some of the very early experimental work, particularly that of Tesla, who seems to have mixed science and showmanship.

Perhaps because RF is unseen, it has also become confused with ionising radiation in the minds of many people. It is essential to distinguish the difference between the two since, with our present state of knowledge, the consequences of exposure to them can confidently be stated as being very different.

Although we cannot see radio waves, most people will, at school or college, have done the classical experiments with magnetic fields and iron filings to demonstrate the patterns of the fields and used an electroscope to demonstrate the presence of electrostatic charge and the force which causes the gold leaf to move.

From these early and rudimentary experiments with static fields it should at least be possible to conceive that such fields are not magical and are very common in any electrical environment.

History of radio transmission

Radio transmission is, relatively speaking, a very new technology which had its beginnings in the theoretical work of Maxwell in the nineteenth century and the experimental work of Hertz, the German physicist, in the last two decades of the nineteenth century. Many others also made contributions, including the development of devices which could detect the presence of radio waves. Whilst the question of who first transmitted radio signals is not without controversy, the subsequent practical development of radio communications systems is attributed to Guglielmo Marconi who was born in Italy in 1874.

His first British patent was taken out in 1896 and covered the use of a spark transmitter. There are many accounts written of the experimental work carried out at various locations on land and on ships during the course of which the range of such equipment was very much increased. By 1921, the thermionic transmitter tube

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2001, Ronald Kitchen. Published by Elsevier Ltd 2001. All rights reserved

became available and made it possible to design transmitters to operate on a range of frequencies. The power output available increased with the development of electronic tubes which could, increasingly, handle higher powers with the aid of air or liquid cooling systems.

Over the years, and stimulated by the needs of the First and Second World Wars, radio transmission has become an established technology which is taken for granted and which, among other things, provides for the broadcasting to our homes of entertainment, news and information of every kind in both the radio and television spheres. The most recent development, resulting in the domestic satellite dish antenna, brings the quasi-optical nature of microwaves to the notice of the consumer.

The use of semiconductor devices (transistors) has become commonplace and as a result the mass and volume of electronic products for a given function is much less than that of their earlier counterparts which used electronic tubes. However, in the high power transmitter field electronic tubes are still the mainstay of transmitters. These use very high voltages, depending on power output. A voltage of 40 kV or more is not unusual for very high power equipments. High power systems such as MF and HF broadcasting systems need considerable provision for cooling the vacuum tubes used and in some cases the resulting heat is transferred to the station heating system!

Semiconductor devices are being used in transmitters of more modest power and also in spaced array radar equipments and do not need high voltages. Semiconductor devices do also have a considerable role in transmitter drives, audio circuits and in control systems. In the latter application, sophisticated logical control circuits are easy to achieve and occupy the smaller volumes attributable to the small size of transistors and integrated circuits.

With the vast increase of terrestrial and satellite broadcasting and communications, and the enormous number of mobile phones now in use, homes, work and recreational places are irradiated by a vast number of electromagnetic signals. Many are intended to operate receiving equipment, most of which are at very low levels because the high sensitivity of receivers does not necessitate large signals. Mobile phones do however communicate both ways and thus incorporate transmitters and receivers. As usage increases there is pressure for the use of more frequencies such that governments now sell licences to use parts of the RF spectrum.

Some radiation is unintentional, resulting from the leakage of energy from devices which have no radiation function, for example, due to inadequate shielding, unblocked apertures in metal cases, and similar shortcomings. Apart from any effects of leakage on people, it also causes interference with other equipment. It is not surprising that the presence of so much electromagnetic interference has caused people to question whether they can be harmed by it.

The word 'wireless' largely passed out of use many years ago. Radio is now the more general term in use, though strangely enough in domestic use it tends not to have the same wide use, mainly being interpreted as meaning sound broadcasting with the term 'television' or 'TV' to describe television picture and sound broadcasting. There are many words used to describe forms of radio system including satellite communications, radar, microwave links, mobile telephones, etc.

Despite the profusion of terms in use to describe the transmission of intelligence by electromagnetic waves, the nature of these waves is basically the same, the variable being the way in which the intelligence (signal) is added. It is therefore convenient to refer to these electromagnetic waves as 'radio waves' and the frequencies of the waves as 'radio frequencies'.

The nature of radio waves

Most readers will be familiar with the fact that an alternating current or voltage which is undistorted has an amplitude which varies with time and reverses direction at each 180°, one cycle taking 360°. This pictorial representation of a current or voltage is referred to generally as a waveform and the description above is that of a sine wave. Waveforms may have other shapes such as square waves, ramps, etc., as will become apparent later.

A sine wave is illustrated in Figure 2.5-1 and is shown with the 'Y' axis denoted arbitrarily in amplitude (A). The term amplitude is used to refer to the magnitude of the voltage or current.

The instantaneous amplitude (amplitude at a specified point in time) can be read from such a diagram and will be found to follow a sine curve, i.e. it is equal to the maximum amplitude (A) multiplied by the sine of the corresponding angle.

Hence at 0° and 180° the instantaneous amplitude is zero. Similarly at 90° and 270° the instantaneous amplitude is at the maximum A but since the sine of 270° is negative the polarity and hence the direction of current flow has reversed. This diagram is basically applicable to any simple AC waveform. One of the factors which distinguishes such waveforms is the time duration of one complete cycle (T) in Figure 2.5-2 and another the frequency (f).

Frequency is simply the number of cycles per unit time and the international convention is 'per second'. The unit is the hertz (Hz) named after the German physicist, 1 hertz corresponding to 1 cycle per second. It follows that the time of 1 cycle in seconds is given by the reciprocal of the frequency in hertz.

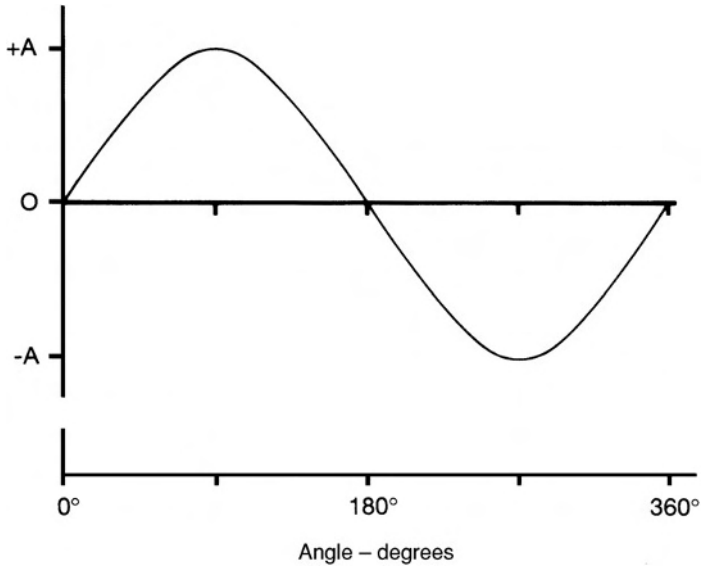


Figure 2.5-1 Sine wave illustration.

The AC mains supply frequency (50 or 60 Hz) is referred to as a low frequency whereas the frequencies used for radio transmission are much higher frequencies. The time T for the duration of a cycle at, for example, 50 Hz is 1/50 s = 20 ms (twenty thousandths of a second) whereas the time for higher frequencies is much shorter as shown in the examples below:

$$T\,(s) = 1/f\,(Hz)$$

Examples:

1. f = 100 kHz T = 10 μs (10^{-5}s).
2. f = 1 MHz T = 1 μs (10^{-6}s).
3. f = 1000 MHz T = 1 ns (10^{-9}s).

For those unfamiliar with these SI prefixes (μ, n, etc.), see Table 2.5-1 which lists those actually used in everyday work.

If the existence of two identical waveforms as shown in Figure 2.5-3 is considered, it is possible for these to be displaced along the time axis so that whilst they are identical in form, the starting points of the cycles may not be identical, i.e. there is a phase difference between the two waveforms.

This may be expressed in angular terms, e.g. 90° phase difference. If two such identical waveforms are exactly in phase and are added, the amplitude of the resultant at any point will be twice that of either waveform alone.

Conversely, if the two waveforms are 180° out of phase the sum will be zero. This becomes relevant when considering radiation surveys and it is necessary to consider the additive possibilities of radio waves reflected from the ground and from metal masses. Obviously additions increase any potential hazards whereas cancellations are less significant in this context since the safety measurement activity is essentially concerned with the highest levels present.

Readers will also be familiar with the idea that a current flowing in a conductor gives rise to a magnetic field around it. When such a current is varying, it gives rise to a similarly changing electric field. Similarly

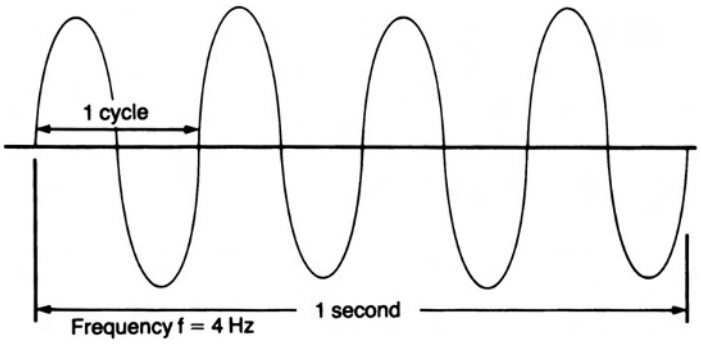


Figure 2.5-2 Sine wave frequency and time relationship.

Table 2.5-1 The most commonly used International System (SI) prefixes.

Symbol	Name	Factor
k	kilo	10^3
M	mega	10^6
G	giga	10^9
T	tera	10^{12}
m	milli	10^{-3}
μ	micro	10^{-6}
n	nano	10^{-9}
P	pico	10^{-12}

a changing electric field will give rise to a magnetic field. Unchanging fields of either kind will not result in the production of the other kind of field. With changing fields the magnetic field and electric field are thus inextricably linked. Hence alternating currents and voltages do, by definition, involve time-varying fields.

It is easy to imagine that from any source of such fields some energy may be unintentionally released (transmitted) into free space, causing interference with receivers or other equipment, without necessarily understanding the phenomenon. This is because such 'interference' has been experienced by most people in their everyday lives. Perhaps the most common example is the motor car ignition system which can also prove to be a rudimentary example of the spark transmitter!

In the case of radio transmitters, however, the whole intention is to transmit RF energy into free space and the antenna used to do so is specifically designed to achieve this objective. If we consider the frequencies discussed above, the very low frequencies, e.g. mains power frequencies, do not give rise to any significant amount of radiation. However, as we increase the frequency then it becomes increasingly possible to radiate electromagnetic waves, given a suitable antenna to act as an efficient 'launcher'.

The electric and magnetic field quantities mentioned above perhaps need a little more elaboration. The electric (E) field at any point is defined as the force acting on a unit positive charge at that point. The magnitude of the electric field is expressed in volt per metre ($V\,m^{-1}$).

The magnetic field at a point is also a force and is defined as the force which would act on an isolated north pole at that point. The classic demonstration of this is that the earth's magnetic field exerts a force on a compass needle, to the great blessing of navigators. The ampere is defined on the basis of the magnetic force exerted when a current flows in a conductor and magnetic field strength is measured in ampere per metre ($A\,m^{-1}$).

Being forces, both quantities are vector quantities having magnitude and direction. The normal Ohm's law equations for power when the voltage and current are in phase (plane wave conditions) can be used in an analogous way and with the same phase qualification to calculate power density.

Plane wave conditions involve the concept of 'free space impedance' which is given by the expression

$$Z_0(\Omega) = \sqrt{(\mu_0/\varepsilon_0)}$$

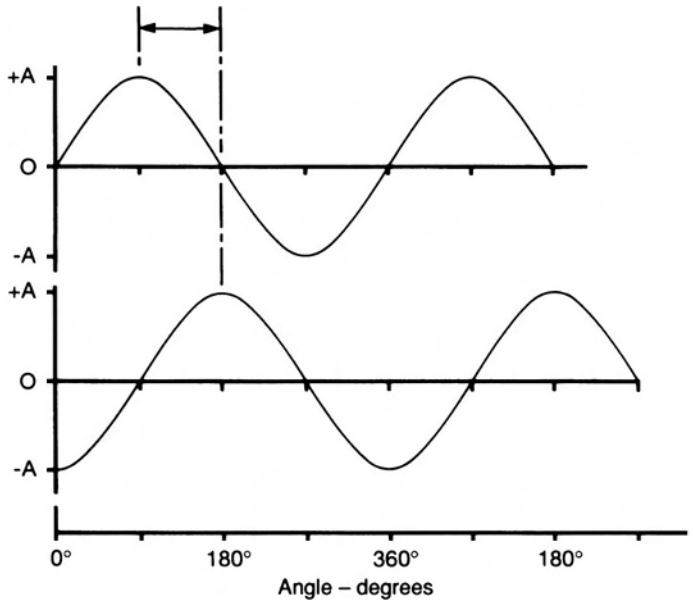


Figure 2.5-3 Phase difference between two waveforms.

where μ_0 is the permeability of free space and ε_0 is the permittivity of free space.

Hence $Z_0 = \sqrt{(4\pi \times 10^{-7}/8.854\ 19 \times 10^{-12})} = 376.7\ \Omega$ (taken as being 377Ω).

For plane wave conditions, $Z_0 = |E|/|H|$ where E and H are field values in $V\,m^{-1}$ and $A\,m^{-1}$ respectively. Hence, under the same conditions, $S\ (W\,m^{-2}) = E^2/Z_0 = H^2 \times Z_0$ where S is the power flux density in $W\,m^{-2}$.

In the USA the most common unit used for S is $mW\,cm^{-2}$ and being the larger unit, is numerically one tenth of the quantity expressed in $W\,m^{-2}$, i.e.

$$1\,mW\ cm^{-2} = 10\,W\ m^{-2}$$

Electromagnetic waves propagated in free space have the electric and magnetic fields perpendicular to each other and to the direction of propagation, as represented in Figure 2.5-4 and are known as transverse electromagnetic waves (TEM waves).

The basic nature of an electromagnetic wave can be physically illustrated by holding two pencils with their unsharpened ends touching and the two pencils being mutually at right angles to each other and held so that one is parallel to the ground and one pointing vertically to represent the planes illustrated in Figure 2.5-4. If now a third pencil is added, mutually at right angles to the other two, it will indicate the direction of propagation as in the figure. The vertical pencil point represents the electric field (vertically polarised wave) and the second pencil the magnetic field.

The plane of polarisation of a wave is, by convention, that of the electric field, i.e. the polarisation in Figure 2.5-4 is vertical. This convention has the advantage that for vertical polarisation the antenna will also be vertical (e.g. a simple rod antenna) and this convention is followed in this book. If the diagram is rotated until the electric field is horizontal then the wave polarisation is horizontal. Apart from this 'linear polarisation', other forms such as circular or elliptical polarisation are also used for specific purposes.

There is another approach to RF radiation whereby the concept of particles (photons) is used to describe the radiated signal. However, for the purposes of this work, the wave concept seems to serve the purpose best and is generally so used.

Frequency and wavelength

Two related characteristics of electromagnetic waves are used as a method of referencing the waves. They are the frequency (already discussed above) and the wavelength. The latter is denoted by the symbol lambda (λ). The relationship between these two characteristics involves consideration of the velocity of propagation of radio waves.

The velocity of propagation of all electromagnetic waves (c) is constant in a given homogenous medium and in free space has a value of $2.997\ 925 \times 10^8\ m\ s^{-1}$ but the approximate figure of $3 \times 10^8\ m\,s^{-1}$ is used in practical calculations. This figure is also used for air but does not apply to propagation in other media. The relationship between frequency and wavelength is

$$c = f\lambda$$

where the wavelength (λ) is the physical length of one cycle of the propagated wave, as shown in Figure 2.5-4.

For electromagnetic waves in free space, where f is in hertz (Hz):

$$\lambda(m) = 3 \times 10^8/f$$

Examples:

1. $f = 200\ kHz \quad \lambda = 1500$ metres
2. $f = 10\ MHz \quad \lambda = 30$ metres

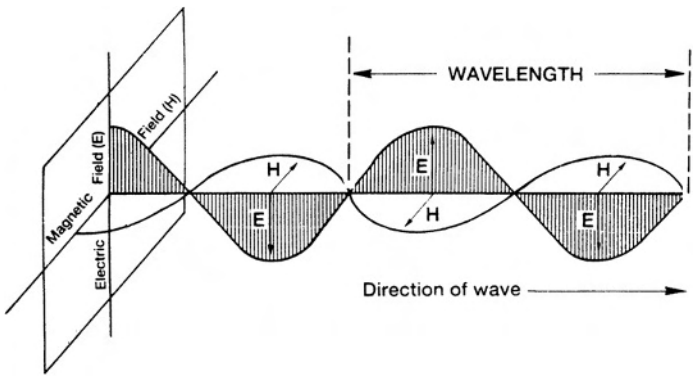


Figure 2.5-4 Representation of a plane wave.

Table 2.5-2 Nominal quarter wave antenna length for a number of frequencies.

Frequency (MHz)	Length – one-quarter wavelength (m/cm)
0.1	750 m
1	75 m
10	7.5 m
100	0.75 m (75 cm)
1000	0.075 m (7.5 cm)
10 000	0.0075 m (0.75 cm)

When f is in MHz, the division simplifies to: $\lambda(m) = 300/f$. This lends itself to easy mental arithmetic! Wavelength is an important parameter in considering antenna systems and propagation since it is a factor in determining the physical dimensions of antennas.

Without going into antenna detail at this stage, some idea of the physical comparison of wavelengths can be obtained from the examples of the length dimension of a $\lambda/4$ (one-quarter wavelength) antenna for a few frequencies shown in Table 2.5-2. Practical antennas will be a little shorter than the theoretical calculations of Table 2.5-2.

Radio waves can therefore be referred to either by the wavelength or the frequency. Domestic receivers may have the scaling in either unit but generally frequency is used, as it is in professional radio work. Wavelength does need to be used when it is involved in determining the physical dimensions of antennas and other devices.

In this book, the range of frequency considered is roughly from 10 kHz to 300 GHz. Table 2.5-3 illustrates the names for the various sub-divisions of the radio spectrum. The term 'microwave', mentioned earlier, does not appear in the listing although with the advent of microwave ovens it has become widely used and misused in the public domain. There is no generally agreed definition but it is often used to apply to frequencies from several hundred MHz upwards.

Table 2.5-3 Frequency band designations.

Frequency	Band code	Band description
300 Hz–3 kHz	ELF	Extra Low Frequency
3 kHz–30 kHz	VLF	Very Low Frequency
30 kHz–300 kHz	LF	Low Frequency
300 kHz–3 MHz	MF	Medium Frequency
3 MHz–30 MHz	HF	High Frequency
30 MHz–300 MHz	VHF	Very High Frequency
300 MHz–3 GHz	UHF	Ultra High Frequency
3 GHz–30 GHz	SHF	Super High Frequency
30 GHz–300 GHz	EHF	Extra High Frequency

Table 2.5-4 Microwave band letters (IEEE).

Frequency (GHz)	Band letter
1–2	L
2–4	S
4–8	C
8–12.5	X
12.5–18	Ku
18–26.5	K
26.5–40	Ka

It should be noted that the term RF is used here across the whole spectrum as a generic term and the term 'microwaves' merely refers to a portion of the RF spectrum.

The abbreviated band identifiers in Table 2.5-3 from VLF to UHF are in frequent use but the abbreviations SHF and EHF are less used, being now increasingly swallowed up in the loose use of the term 'microwaves'. In addition, there is a more specific classification system for bands in the upper UHF onwards.

This is given in Table 2.5-4 on the basis of the IEEE listings. It has to be said that different versions of these band classifications are in use across the world and in textbooks so that reference to frequency is perhaps the only safe way of avoiding ambiguities. The presentation of the different possible classifications tends to confuse rather than enlighten.

Conveying intelligence by radio waves

When a wave of a given frequency is radiated continuously, i.e. a continuous series of sine waves, no intelligence is conveyed and the signal is called a 'carrier'. This mode of transmission is known as continuous wave (CW). Nothing can be heard unless there is a local oscillator to 'beat' with the carrier and produce a note at the difference frequency. This is referred to as heterodyning.

If the carrier is switched on and off in accordance with some kind of code, e.g. morse code, then this intelligence can be interpreted. More generally, for broadcasting the

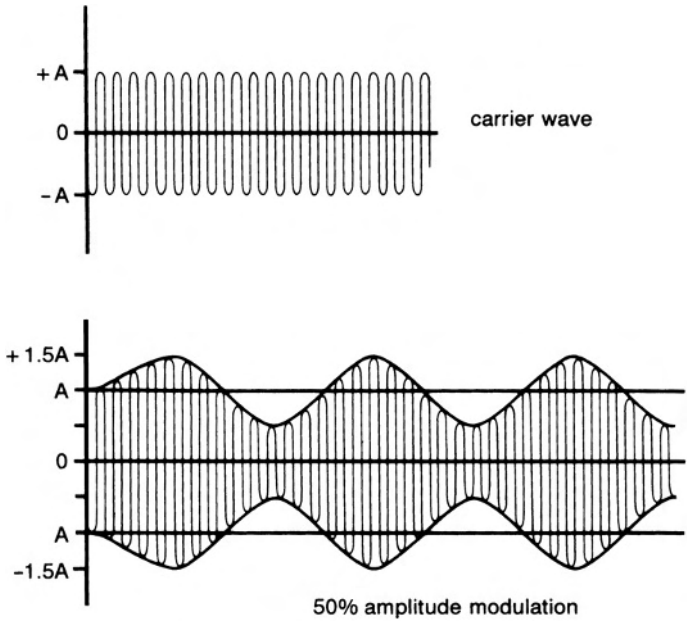


Figure 2.5-5 RF signal, amplitude modulated by a low frequency signal.

intelligence may be speech, music, and television pictures. Other professional work includes voice and data transmission by a variety of methods, radar transmitters transmit RF signals in a series of pulses and so on.

The process of sending intelligence is referred to as the process of modulation and the technical methods of doing so are wide ranging and outside the scope of this book. It is however useful to illustrate the general nature of amplitude modulation which has some significance when carrying out radiation measurements, and also to illustrate the principle of pulse transmission.

Figure 2.5-5 illustrates the waveform of a carrier and of the same carrier with 50% modulation applied in the form of a simple audio frequency sine wave. It can be seen that the peak instantaneous amplitude of the 50% modulated wave is against A for the carrier. Clearly the total power is greater when the carrier is modulated and hence any field measurements made will need to be related to the modulation state. For amplitude modulation, Figure 2.5-6 shows the relationship between sine wave amplitude modulation depth versus transmitted RF power and RF current.

Figure 2.5-7 shows a pulse transmission where the carrier is transmitted for time t_p (the pulse duration) and with a pulse repetition rate of n Hz (pulses per second). It is, of course, not possible to show this to scale since there

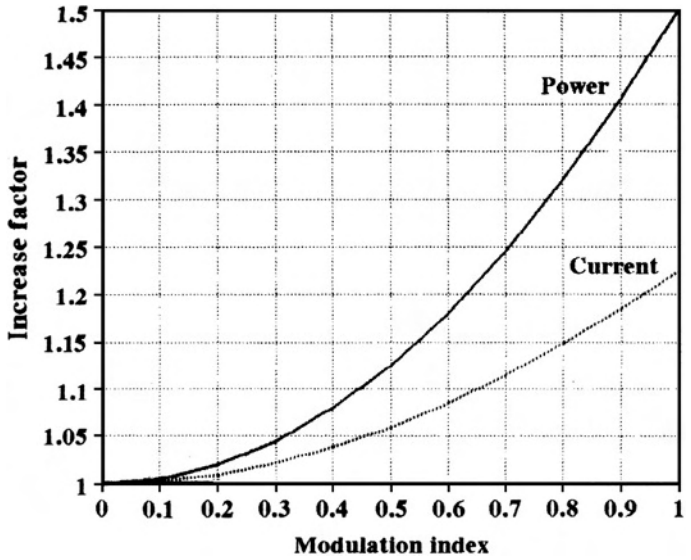


Figure 2.5-6 Sine wave amplitude modulation depth versus RF power and RF current.

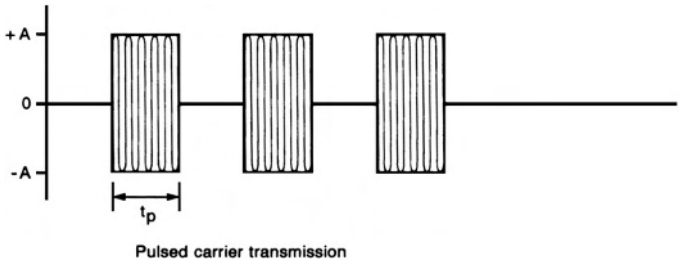


Figure 2.5-7 Pulse modulation.

will be too many cycles of carrier in each pulse to actually illustrate them. For example a radar working at 1 GHz and with 2 μs pulses will have 2000 cycles of carrier in each pulse.

There are many other methods of modulation and transmission which can be applied to radio equipment and which cannot be covered here but which need to be known to those doing safety surveys.

Much power is wasted in amplitude modulation and various other forms which reduce the waste are widely in use such as single sideband (the two sidebands in AM contain identical intelligence) with a reduced carrier power, double sideband where the two sidebands carry different intelligence and again the carrier is reduced, etc.

The recent development of digital radio and television will likely become widespread over the next few years.

Ionising and non-ionising radiations

Confusion between these two forms of radiation amongst the public has been mentioned earlier. There is also a surprising amount of misunderstanding amongst electronics and radio engineers about the distinction between these two forms of radiation even amongst newly qualified graduate engineers, so that RF radiation is sometimes considered to be the same as ionising radiation.

Ionising radiation, by definition, is radiation capable of ejecting electrons from atoms and molecules with the resultant production of harmful free radicals. There is a minimum quantum energy below which this disruption cannot take place. Since the human body is largely water, the water molecule is used to define this minimum level.

Different reference sources give varying figures for this between 12 eV and 35 eV. The actual value does not matter for the purposes of this comparison. An energy of 12 eV corresponds to a wavelength of 1.03×10^{-7} metres (103 nm) which can be seen from Figure 2.5-8 lies just above the ultraviolet (UVc) spectrum. The highest RF frequency used in standards for RF safety is 300 GHz which corresponds to a wavelength of 10^{-3} metres and lies in the EHF band of the radio frequency spectrum. If the calculation is done the other way round, 300 GHz corresponds to an energy of 0.00125 eV which, from the foregoing, is too small by about four orders to cause ionisation.

However, in radio transmitters using very high supply voltages, ionising radiation in the form of X-rays are produced. It should be clear that this ionising radiation is not inherent in the RF energy but rather that both forms of radiation can co-exist inside equipment and the RF engineer or technician needs to be aware of the hazards involved. It is also the case that ionising radiation is, in most countries, subject to definitive legal provisions due to its hazardous nature.

Explanation of terms used

In this section those terms and units which are most frequently used in dealing with RF radiation are explained. The more formal definitions may be found in reference books. Other more specialised terms are introduced in the text as appropriate.

1 Transverse electromagnetic mode wave (TEM)

An electromagnetic wave in which the electric and magnetic fields are both perpendicular to each other and to the direction of propagation (see Figure 2.5-4).

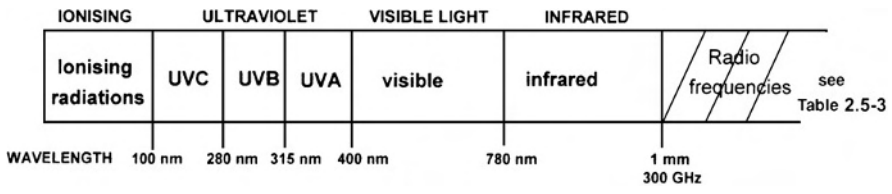


Figure 2.5-8 Radiation wavelengths relative to 300 GHz (WHO).

2 Power

The rate of doing work in joules per second. The unit is the watt (W) which corresponds to 1 $J\,s^{-1}$. Sources of RF energy are rated in watts. Both the kilowatt (kW) and the megawatt (MW) are common in radio work, the latter typically for very high power equipment such as radar equipment.

3 Mean power

The r.m.s. power supplied or generated averaged over a period of time which is long compared with the longest period of any modulation component.

4 Power flux density (power density)

Power flow per unit area through a surface normal to the direction of that flow, usually expressed in watt per square metre ($W\,m^{-2}$). However it is also often quoted in $mW\,cm^{-2}$. The use of hybrids such as $W\,cm^{-2}$ are best avoided except where really necessary, since they can cause confusion. Here the shorter form in common use 'power density' is used hereafter because of the frequent occurrence in text. All references to power density, electric field and magnetic field are to r.m.s. values, unless otherwise stated, in common with the practice in RF safety standards.

5 Energy density

This is, strictly, related to volume ($J\,m^{-3}$) but is almost universally used in radiation protection work as the product of power density and time and expressed either in units of watt-hour per square metre ($W\,h\,m^{-2}$) or joule per square metre ($J\,m^{-2}$). 1 J = 1 W s. It is sometimes used to express a total energy limit, for example, 'not more than 5 $W\,h\,m^{-2}$ in a six minute period'.

In terms of the energy in a volume, e.g. $J\,cm^{-3}$, the definition relates to the energy in a minute volume divided by that volume. With a power density of 10 $W\,m^{-2}$ the energy in a cubic centimetre of air is 0.033 picojoules.

6 Electric field strength (E) at a point

A vector quantity defined as the force acting on a unit positive charge at that point. It is expressed in volt per metre.

7 Magnetic field strength (H) at a point

A vector quantity defined as the force acting on an isolated north pole at that point. It is expressed in ampere per metre.

8 Specific absorption rate

The rate of absorption of RF energy in a substance, normally human tissue, expressed in watt per unit mass, e.g. watt per kilogram. If the substance is not human tissue, it should be specified. Note that a specific absorption rate (SAR) limit may be expressed in this standard form but be limited to a maximum mass of tissue e.g. 10 $W\,kg^{-1}$ (10 g) should be interpreted as an SAR of 10 $W\,kg^{-1}$ in any 10 g of tissue.

9 Frequency

The number of cycles of an alternating current per unit time where the international period is 1 second. The unit is the hertz. 1 Hz = 1 cycle per second.

10 Pulse repetition frequency (p.r.f.)

In a system which uses recurrent pulses, the number of pulses occurring per unit time. The unit is the hertz (Hz).

11 Peak pulse power density

In pulsed systems such as radar equipment the term 'peak pulse power' is used when what is actually meant is the r.m.s. power in the pulse (see Figure 2.5-7). This should not be confused with instantaneous peak power.

12 Pulse duty factor

Where t_p is the pulse duration in seconds and n is the pulse repetition rate in Hz, then the duty factor $DF = t_p n$ and has a value less than 1.

For example, if $t_p = 2\ \mu s$ and n = 500 Hz, then:

$$DF = 500 \times 2 \times 10^{-6} = 0.001$$

Many people find it easier to work with the reciprocal, in this case 1/0.001 = 1000.

The relationship between peak pulse power density (S_{pk}) and the mean power density (S_{mean}) in a pulsed system is:

$$S_{pk} = S_{mean}/DF$$

or, if using the reciprocal of DF, specific absorbtion rate

$$S_{mean} \times 1/DF = S_{pk}$$

Note that although pulse transmission often seems to be uniquely linked to 'radar', pulse transmission is widely used and radar is just one application. Note also the high values of S_{pk} which are possible, depending on the duty factor.

13 Antenna (aerial)

The generally used term for any type of device intended to radiate or receive RF energy. These range from simple

wires and rods to arrays (of which the television antenna is an example) to large microwave parabolic, elliptical and rectangular aperture systems.

Some antennas are dedicated to reception or transmission whilst others do both. To most people the terms antenna and aerial are synonymous. The English plural is normally used for antenna.

14 Antenna, isotropic

A hypothetical, idealised, antenna which radiates (or receives) equally in all directions. The isotropic antenna is not realisable but is a valuable concept for comparison purposes.

15 Directive gain of an antenna

The ratio of the field strength at a point in the direction of maximum radiation to that which would be obtained at the same point from an isotropic antenna, both antennas radiating the same total power.

16 Antenna beamwidth

The angular width of the major lobe of the antenna radiation pattern in a specified plane. The usual criterion for beamwidth is to measure between points either side of the beam axis where the power density has fallen to half (3 dB down) of that on the axis. This is usually referred to as the '3 dB beamwidth'.

17 Equivalent radiated power (ERP)

The product of the power into the antenna and the gain referred to a dipole. It is often used to specify the power of UHF/VHF broadcast transmitters.

18 Equivalent isotropic radiated power (EIRP)

The product of the power into the antenna and the gain referred to an isotropic antenna.

19 RF machines and RF plant

RF energy is now increasingly used to undertake manufacturing operations which use heating and these terms are used here to refer to machines generally. In practice they have functional names, e.g. plastic bag sealer, plastic welder, etc. Their significance is that they use an RF generator which, in terms of safety, needs the same consideration as any other RF generator.

Use of the decibel

Whilst most people trained in electrical and electronic engineering will have covered this topic, experience in running RF radiation safety courses shows that whilst many people work regularly with decibels, a surprising number never have occasion to do so and it is often necessary to do a refresher session in this topic.

The bel and the decibel (one tenth of a bel) were originally used to compare sound intensities and are currently used in safety legislation to limit the exposure of people to intense sounds in the workplace. Some safety officers will be familiar with this method of noise control.

In radio work, the decibel is used to compare powers, voltages and currents. The decibel is a dimensionless number representing a ratio based on common logarithms. However, usage is such that the ratio is often referenced to a value of a quantity so that it can be converted to a specific value of that quantity. This is a practice of convenience which has developed, so it is best to start with the basic role of the decibel as a dimensionless number. The bel itself is not normally used in radio work.

Decibels and power

If we wish to compare two powers, P_1 and P_2 , then we can do so by dividing one by the other. The resulting ratio is P_1/P_2 and is a pure number. To express this in decibels the form is:

$$\text{Ratio (dB)} = 10 \log(P_1/P_2)$$

If $P_1 = 1600$ W and $P_2 = 2$W then the simple ratio is:

800

The ratio in decibels is $10 \log 800 = 29.03$ dB.

Since the decibel is based on logarithms, a number of simplifications follow. The basic rules for ratios which are pure numbers are therefore:

1. Multiplying numbers merely requires the addition of the decibel values.
2. Dividing numbers requires the subtraction of one decibel value from the other.

As powers can be in kilowatts or megawatts, it can be seen that the arithmetic involved is much simpler, especially as gains can also involve inconveniently large numbers, e.g. 69 dB gain=7 943 282.

To convert decibel values back to plain ratios we reverse the process:

For 29.03 dB in the earlier example, the ratio is given by:

antilog $(29.03/10) = 10^{2.903} = 800$ as in the first calculation

Decibels and voltage

Since power can be expressed as V^2/R, then the ratio of two such expressions, where V_1 and V_2 are the two voltages and R_1 and R_2 the corresponding resistances, is:

$$(V_1^2 R_2)/V_2^2 R_1$$

If $R_1 = R_2$, then the ratio is now:

$$(V_1)^2/(V_2)^2$$

and $$\text{dB} = 10 \log (V_1)^2/(V_2)^2$$

$$= 20 \log V_1/V_2$$

Hence for voltage ratios, the formula for conversion is:

$$\text{Voltage ratio (dB)} = 20 \log \text{voltage ratio}$$

Referencing ratios

So far we have considered dimensionless quantities where the rules for handling the resultant dB values are those related to the use of logarithms generally. It is possible to reference ratios to any quantity, a common one being the milliwatt. The usual reference to this is dBm rather than the expected dBm W. Table 2.5-5 shows some decibel values referenced to 1 watt (dB W) for powers greater and smaller than the reference value.

Table 2.5-5 Table of power (watts) versus decibel value relative to 1 watt.

Watts	dBW
1000	+30
100	+20
10	+10
1	0
0.1	−10
0.01	−20
0.001	−30

This is a very convenient way of handling power in calculations.

To convert back to watts power, the process is as before except that the ratio obtained is multiplied by the reference value:

$$1000 \text{ W} = 30 \text{ dB W}$$

To reverse this:

$$\text{Power (W)} = \text{antilog } (30/10) \times \text{reference value 1 W}$$
$$= 1000 \times 1(\text{watts})$$

When the reference value is unity, in this case 1 watt, the last multiplication is academic.

This page intentionally left blank

Section **Three**

Network communications

This page intentionally left blank

Chapter 3.1

3.1

Data and voice traffic

Juanita Ellis, Charles Pursell and Joy Rahman

In the past, voice and data networks had been kept separate. Legacy network technologies simply could not meet the diverse performance requirements of both voice and data. Advances in networking technology, including fast Ethernet, wire-speed switching, and Policy-Based Quality of Service (QoS) management, have made it possible to build converged voice and data networks. Converged networks enable a new generation of integrated voice/data applications. For example, with converged networks, users of web-based e-commerce applications can view product information while talking with customer service agents in a call center, through a single network connection. The focus of most converged network strategies is Voice over IP (VoIP). VoIP refers to the transmission of telephone conversations over a packet-switched IP network. This IP network could be as small as a single subnet, private LAN, or as large as the public Internet. With VoIP on the LAN, telephone conversations are converted to a stream of IP packets and sent over an Ethernet network. This network is usually restricted to a building or campus. As VoIP technology matures, new conversion methods may emerge. Regardless of the method that is used to convert VoIP traffic for LANs, VoIP traffic will always traverse the LAN as a stream of IP packets. One of the key challenges in implementing VoIP is to design and build an IP-based network that meets QoS requirements and is comparable in performance to conventional circuit-switched telephone networks. The high latency forwarding and best-effort delivery provided by traditional software-based routers is generally not acceptable for streaming traffic such as VoIP because it does not provide maximum latency guarantees or minimum bandwidth guarantees. To understand the core architecture of converged network, we must understand the behavior of voice and data traffic and how networks perform on different network traffic.

Packet switching versus circuit switching

There are primarily two types of networks, circuit switched and packet switched (Table 3.1-1). A good example of circuit switching is the telephone system. Circuit switching is a technique in which a system seeks out the physical "copper" path from the caller's telephone to the receiver's telephone. There is a need to set up an end-to-end path before any data can be sent. This is why it can take several seconds between the end of dialing and the start of ringing for international calls. During this interval, the telephone system is actually hunting for a copper path, and the call request signal must propagate all the way to the destination and be acknowledged. No one else can use the physical path at the same time.

On the other hand, packet switching is a type of network where small units of data are routed through a network based on the destination address in each packet. Compared to circuit switching, which statically reserves the required bandwidth in advance, packet switching acquires and releases bandwidth as and when it is needed. This type of communication between sender and receiver is known as connectionless. Circuit switching is completely transparent. The sender and receiver can use any bit rate, format, or framing method they desire. As for packet switching, the carrier needs to determine these basic parameters first before transmission can occur.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2003 Elsevier Inc. All rights reserved

The Internet is based largely on packet switching, and the Net is basically a huge connectionless network joined together. By transmitting data in packets, the same data path can be shared among many users in the network.

Circuit-switched networks are based on Time Division Multiplexing (TDM), in which numerous signals are combined for transmission on a single communications line or channel. Once a connection is made, a connection is established and remains throughout the session. On the other hand, in packet-switched networks relatively small units of data called packets are routed through a network based on the destination address contained within each packet. Breaking communication down into packets allows the same data path to be shared among many users in the network.

In converged networking, packet-switching technology is used to carry the information throughout the network. Converged network supports all of the user's traffic types with packet-based protocols such as ATM, Frame Relay, or IP. Of these, the dominant protocol in the access, edge, and core of the network is the Internet Protocol (IP). A managed IP infrastructure for voice, video, and data transmission is a concept that has begun to revolutionize the industry. It leverages the Internet so that enhanced services can be offered at lower costs and it motivates the development of a family of dynamic, next-generation, real-time applications. Voice, video, and data traffic have significantly different characteristics and are difficult to reconcile within a single network. Data traffic tends to be bursty, consuming large volumes of bandwidth for occasional, short intervals, whereas voice traffic is predictable and requires a steady, low-delay, transmission path from end to end. Video is a hybrid of both.

- Most computer networks currently use "packet switching" rather than "circuit switching"

Circuit switching

- Circuit switching (see Figure 3.1-1): local exchanges connect to remote exchanges
- Each caller may make one call at a time
- Line is unusable by others until call terminated

Example: POTS (plain old telephone system)

Packet switching

- Technique used with digital connections to allow multiple calls to exist on same circuit
- Data broken into small blocks ("packets")
- Packet includes extra information in Header (serial number, destination address, etc.)
- Each packet routed to its destination individually
- Packets re-assembled into original message when they reach destination

Advantages of packet switching

- Network supports many connections simultaneously
- Short messages not delayed by long messages
- More efficient than circuit switching

Disadvantages

- Performance drops when many users share the same network

As far as communication networks are concerned, the notion of switching began with circuit switching. The telephone (or telegraph) company provided an electrical path that allowed my instrument to connect to yours, perhaps with an operator plugging a connector into a jack. The telephone company first combined (or multiplexed) multiple calls on a single physical circuit using Frequency Division Multiplexing (FDM). You can think of these discrete calling paths as the first virtual circuits in the sense that there was no longer a one-to-one ratio between phone calls and wires (or insulators on the utility pole).

Table 3.1-1 A comparison table on circuit switching vs packet switching.

	Circuit switching	Packet switching
Dedicated "copper" path	Yes	No
Bandwidth available	Fixed	Dynamic
Potentially wasted bandwidth	Yes	No
Store-and-forward transmission	No	Yes
Each packet follows the same route	Yes	No
Call setup	Required	Not required
When can congestion occur?	At setup time	On every packet
Charging	Per minute	Per packet

Frame switching

FDM turned out to be insufficiently scalable for the demands of telephony. So, in the early 1960s the phone companies began digitizing voice signals and multiplexing them in the time domain using TDM.

For example, with TDM a T1 line interleaves 24 phone calls among successive time slots within 1 frame,



Figure 3.1-1 Circuit switching.

which consists of 193 bits. Bit 1 through bit 8 are dedicated to channel 1, bit 9 through bit 16 are dedicated to channel 2, and so on, until bit 185 through bit 192 are dedicated to channel 24. The 193rd framing bit is used to synchronize the system. The interleaving process repeats 8000 times each second. (Note that 193 bits per frame times 8000 frames per second equals 1.544 Mbits/sec, which is the throughput rate of a T1 line.)

Today, TDM phone call switching is circuit oriented, though the devices that perform the switching function on digital circuits bear little resemblance to the mechanical switches that once selected paths made up of solitary electrical circuits.

One disadvantage to TDM—whether you're making a call with a TDM system or leasing a full-time digital line—is that your cost will be the same whether you fill every time slot with data or transmit nothing (for our purposes, let's consider voice to be just another form of data). As we know, the data transmissions required for many applications are bursty, with intervals of high demand and no demand distributed almost randomly. Thus, TDM-based networks (or any circuit-oriented networks with hard resource allocations) are likely to be inefficient or otherwise not wholly suitable for data traffic.

Ultimately, dedicated circuits are a high-cost form of connectivity, and setting up a circuit via a switching system designed for voice communications results in long circuit-initiation times, as well as high cost.

Packet switching

In the late 1960s the notion of packet switching was developed (see Figure 3.1-2). The first commercial outgrowth of this technology was the X.25 network. In spite of outbreaks of Internet fever in many countries, it is still heavily relied on in much of the world. Packets on an X.25 network aren't slotted rigidly in the way that circuits in the TDM system are. Instead, the packets are created and transmitted as needed. Therefore, X.25 service can be priced by the packet or by the byte rather than by connection time or as a full-time circuit; you're using network capacity only when you're sending or receiving data.

Despite this freer form of multiplexing, X.25 networks are still connection oriented, and a session between two nodes still requires a virtual circuit—the virtual circuit has just been unbundled from a fixed time slot.

Each packet in an X.25 network has a Logical Channel Number, or LCN. When a packet comes into a switch, the switch looks up the LCN to decide which port to send the packet out of. The path through the network of packet switches is defined in advance for Permanent Virtual Circuits (PVCs) and established on the fly for Switched Virtual Circuits (SVCs). With SVCs, call setup is required before data transmission can take place. (Incidentally, the protocol data units at Layer 3 are known as packets, while the protocol data units at Layer 2 are called frames, at least when writers are precise. The X.25 protocols include Layer 3 functions, while frame relay, which is essentially X.25 with error correction and flow control removed, remains at Layer 2.)

Connectionless switching

Local Area Network (LAN), which people began to develop in the 1970s, is a form of connectionless communication. When a Layer 2 bridge connects LANs, a form of frame switching takes place. When a Layer

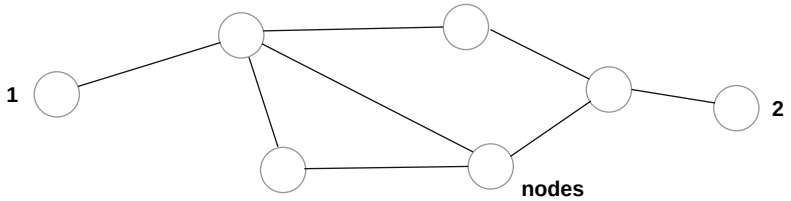


Figure 3.1-2 Packet switching.

3 router connects LANs, a form of packet switching takes place. IP and its precursors were the first wide-area connectionless protocols used extensively. Each packet includes its source and destination addresses and moves independently through the network. With these connectionless switching systems, for the first time there was no advance setup of the path needed. There was also no advance agreement on the part of the intermediate stages committing to a specific level of service. There was no state maintained on the network's packet switches, which have no notion of circuits, paths, flows, or any other end-to-end connection.

There's an economic and fault-tolerance factor you should keep in mind when considering the router-using flavor of packet switching. The economic issue revolves around the relative costs of computer cycles and communications circuits. If processing power is cheap, it's not unreasonable to figure out the routes every packet should take in order to make the best use of expensive circuits. If circuits are cheap, you may prefer to set up a mesh of connections using relatively dumb circuit switches, rather than dedicate a bunch of high-powered, special-purpose computers to routing packets.

The fault-tolerance issue (often expressed as the overheated claim that the Internet was designed to survive a nuclear war) is that connectionless networks are more resilient than connection-oriented ones. If a link or a router goes down, the overall system is designed to find routes around the problem. A broken connection-oriented network, designed to remember how to handle each circuit rather than to solve a routing problem for each packet, will likely need manual reconfiguration if it consists of PVCs. Furthermore, if it consists of SVCs, it will at least need to perform a call setup.

While most of the voice links are circuit switched, there does exist some packet switching of voice data in the telephone network. Beyond the local loop, a large part of the signal and control network is packet switched like IP, though the signaling networks generally use a different suite of protocols. Circuit switching and packet switching address the same problem. A lot of data comes into a communications computer (switch) on a set of incoming channels and must be directed to the right outgoing channels. The key difference is how resources are reserved on the channels and within the communications computer.

Where the traditional network deployed today is designed to carry *circuits*—fixed channels across which communications between telephones or other devices are conducted—the telecommunications industry is now preparing for a big technology changeover, what's being called a "next-generation network" based on so-called data packets.

Instead of setting up an explicit channel for each conversation or data flow, all the traffic types—data, video, voice, etc.—will travel together across the network links. Each data flow is broken up into packets (variable-length chunks of data), each carrying its own address, with the packets being forwarded by the network to their ultimate destination based on the information they carry with them.

Most importantly, the new network will separate services and infrastructure. In effect, the network is moving from being the equivalent of a rail system, where design fundamentals dictate the specifics of the train (such as the wheel gauge and carriage shape), toward the equivalent of a road system, where any type of vehicle ranging from a motorcycle to a truck can travel on a single, simple infrastructure. Like the road system (and unlike the rail), traffic is granted its own autonomy. Instead of working to a strict, centrally controlled timetable and signaling system, the packets simply set out and steer themselves toward their destinations.

The aptly-named "superhighway" therefore has a series of advantages. It will not only be able to run different sorts of traffic at high speed, but it will be designed in such a way that it can accommodate further change. Unlike the real-world highway, the telecom highway can mandate faster and faster speeds and a vast increase in the number of lanes running between any two points. Like the real-world highway, its strength is its simplicity. The self-addressed packet standard is the equivalent of tarmac. The telecom equivalent of new tires and whole new vehicles can be designed and introduced without detailed reference to the underlying network infrastructure, or roadway. Packet-based networks are not without their problems, however. The main issue for voice traffic is that it has to arrive at its destination not only almost instantaneously but also in coherent chunks to be comprehensible. Connectionless packet routing using the IP is unable to secure guaranteed quality of transmission. Everything depends on how much other traffic there is on the network at any one time. But there are *solutions*. IP is a real break with the past. For the first time the telecom environment will look more like the computer environment, where a standard operating system will enable both equipment vendors and applications providers to innovate to provide value-added services. The result will be good news for users, who will see costs (especially for voice) rapidly diminish. At the same time, the new environment should stimulate a faster innovation cycle with new residential and business services being brought to market in months rather than years or decades.

Data traffic characteristics

In the early days of data communications, all networks were circuit-switched, and many still are. More recently,

for networks that cover wide areas, the emphasis has shifted to packet switching, simply because it permits the interconnection of far more nodes into a single network. With packet switching, fewer communication channels are required (because channels are shared by many users), and interconnection of networks is much easier to accomplish.

The data layering of communications gives rise to situations where the lower layers of a network are connectionless, but the higher layers establish a connection; the opposite, where the lower layers establish a connection and the upper layers do not, can also occur.

Broadcast channels

To understand data network we have to know the concept of communications channels of different types of networks. Broadcast channels are, as the name implies, similar to short-wave radio conversations. A node broadcasts on the common channel and all other nodes connected to the channel listen to the message (see Figure 3.1-3).

The channel might consist of electromagnetic emissions (that is, radio), or a wire or cable to which all nodes are attached. As with users of CB radio channels, more than one node might begin to transmit at the same time. When that happens, each garbles the other's message. They all must stop and then start over again in such a way that subsequent collisions are not likely to happen. Alternatively, transmissions can be scheduled so that more than one transmission never takes place at the same time.

Data networks can be classified according to the area over which they extend. A LAN consists of a few up to several hundred nodes, but will typically be confined to a few buildings within a few thousand meters of one another. It can consist of subnetworks linked together in certain ways to form the larger, but still local, network. A subnetwork is a portion of a network in which all of the nodes are directly connected; for example, all of the nodes may be connected by one piece of wire.

Metropolitan area networks (MANs) continue to evolve and will be developed primarily by data carriers in response to the demand to interconnect LANs across a metropolitan area. For example, a university might interconnect its campuses. Wide area networks (WANs) are often interconnected LANs or MANs. They can be homogeneous, but are often heterogeneous, that is, interconnecting LANs or MANs that have been built using different technologies. A WAN can span campuses, cities, states, or even continents. Typically, only one node on each LAN or MAN, called a gateway, connects to the WAN. Other nodes communicate with the WAN via the gateway.

Data communications involve the transfer of data between computer programs. Just as humans must share a common language in order to communicate, the programs must have a common protocol. The protocol simply defines the format and meaning of the data that the programs interchange.

An example of a very simple protocol might be the following: Suppose two programs connected by a communications channel need to exchange messages that vary in length. The protocol might specify that the first three characters of each message be numeric characters giving the length of the message itself (in decimal, not counting the first three characters). For the message "HELLO WORLD" (containing 11 bytes including the space), the sending program would transmit the characters "O11 HELLO WORLD." The receiving program would accept the first three characters and then, knowing the length of the message, would expect 11 more characters (see Figure 3.1-4).

If the receiving program got a message that had something other than numeric in the first four characters, it would consider that an error. If the sending program

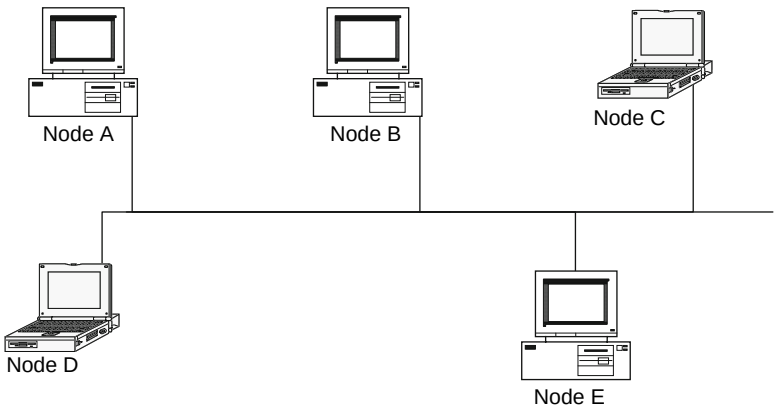


Figure 3.1-3 Sample of broadcast channel example.

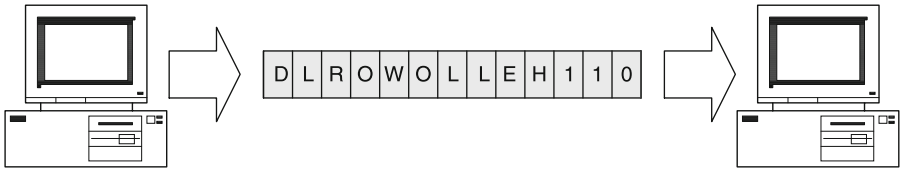


Figure 3.1-4 Computer protocol example.

stopped sending before all of the characters were received, that would also be considered an error.

Early data communications programs were monolithic—that is, a single large program provided many services and communicated with other programs with a single "low-level" protocol (see Figure 3.1-5). This did not work well. The programs were hard to change and the protocols would not work well as new technology came along. These problems led to the development of layering concepts.

A layer is simply a program or set of programs that provides services to the next higher layer and uses services of the next lower layer. A program that resides at the highest layer will typically provide many sophisticated services to the user, but most of these services are actually implemented, directly and indirectly, by the lower layers.

Because a program provides services only to the layer above it and uses services only of the layer below it, a change to any given layer will affect only the layer above it. Layering breaks a single monolithic program into parts that are isolated from one another, making the program easier to write and to change. Layering does, however, extract a performance penalty. There is some overhead associated with moving data through multiple layers.

Protocol layers

Layering applies to protocols as well as to services. In a system that has a layered architecture, a process communicates only with its peer processes on the same layer; otherwise, as with services, a change to one layer would affect many other layers. Peers communicate with a common protocol, appropriate to the services they provide. Each level may require a protocol that is different. So layers of processes have corresponding layers of protocols.

This concept of layering gives rise, in its simplest form of two layers of services and two layers of protocols, to the service/protocol model shown in Figure 3.1-6. Learning this basic model is essential to understanding the remainder of the material in this course.

Layered communications systems

The simplest service/protocol model has two layers, but the ideas it illustrates can be extended to an arbitrary number of layers. In fact, real data communications systems typically involve from three to seven layers of services and protocols.

Peer communications

When two peer programs happen to be running on the same machine, communication will be through the lowest layer in the stack.

Levels of abstraction

The layers of data communications services can be viewed in a general way. Programs at the lowest layer provide services related to the simplest, most "concrete" form of data: streams of bits. Programs at the highest layer provide services related to the most complex or "abstract" forms of data: data that are ready to be displayed to humans, application program data structures, or, in the case of programs using more advanced object-oriented programming techniques, objects. Programs at the intermediate layers transform the data from the simple to the complex forms, and vice versa. This transformation is sometimes referred to as "abstraction," and you might hear the phrase "levels of abstraction," referring to movement from the concrete to the abstract, that is, from the lowest level to the highest.

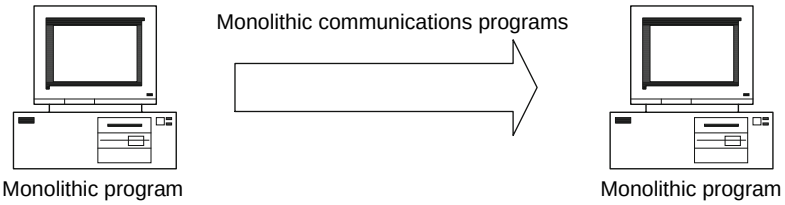


Figure 3.1-5 Early data communications programs.

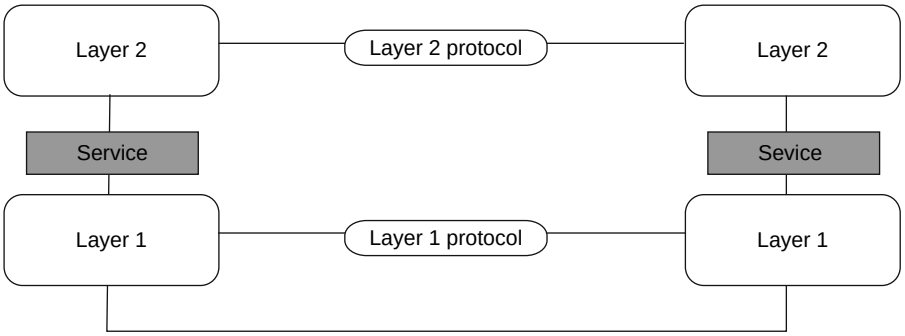


Figure 3.1-6 Protocol layering.
The diagram illustrates the concept of layering. In this example Layer 1 provides a "service" to the layer above, Layer 2. An example of a service would be transmission of bits across a physical media such as an RS232 cable. Layering breaks a large program into a series of smaller and more manageable tasks. Layering also isolates one layer from another layer.

Layering and routing

One reason that protocols vary from layer to layer is because the services offered by each layer to successive layers progress from concrete, bit-oriented services to more abstract, higher-level services for data objects. But protocols differ for another important reason: The lower layers communicate across a single link, while the upper layers must communicate with peers to which the link is indirect, that is, through other nodes.

The bit-stream-oriented protocol at the lowest level supports simple streams of bits flowing between two points. Because the only language understood at this level is "0" and "1," the protocol recognizes only peers to which it is directly connected by the physical link across which the bits flow. It simply has no way of addressing anyone else, and that's not its job anyway.

If communications between peers at all levels were bound by this restriction, communication possibilities would be extremely limited. But a program at a higher level can invoke services from more than one process on the lower level and thus from more than one physical link. Note that the two lowest-level protocols do not need to be identical protocols; in fact, often they will not be. Because higher levels can pass more complex messages, they can include routing information, making it possible for data to flow across more than one communication link, and making large networks possible. So, an important service provided by lower-level protocols in any of the networking systems we will study (which are all layered) is to handle the routing of messages between computers that are not connected by a physical link.

In a system with a layered architecture, programs at a high level deal with more complicated data but don't have to deal with implementation details and routine tasks. For example, in a network the routing of messages is done by lower levels, so upper-level programs need not be aware of it. The data flow only through the highest level that is concerned with handling them. So when a message flows through an intermediate node the data will flow only through the lower-level processes involved with their transmission.

Protocol stacks

The layering of programs gives rise to layers of protocols, often called "protocol stacks" or "protocol suites," that define a communications standard. Anything or anybody that communicates must share a common protocol with that with which it communicates. Programs that communicate with one another using the same protocol are called peers. Peer programs at each layer provide service to the next higher layer, and use the services of the next lower layer to perform their assigned function. As with services, the lowest-level protocol deals with bits; the highest-level protocol deals with complete data structures or objects that are ready to be used by an application program or displayed to a user.

Data networks

A data network connects nodes, some of which are hosts to which terminal nodes attach, in two different ways: point-to-point and broadcast. Point-to-point networks fall into two classes: circuit-switched networks, in which a connection is formed between the nodes, as in a telephone network; and packet-switched or connectionless networks, in which packets of data or datagrams are passed from node to node until they reach their destination, like telegrams. All of the nodes in a broadcast network share the same channel, and the network protocol controls access to the channel and avoids or recovers from the collisions that occur when more than one node tries to use the channel at the same time. LANs extend across a single site and consist of one or more subnets, which are usually, but not necessarily, homogeneous. MANs, the metropolitan area networks

now evolving, will be developed by data carriers to connect LANs in the same city. WANs are wide area networks, often heterogeneous, which cover many sites, spanning large corporations and sometimes continents. Five common topologies exist for point-to-point networks: star, ring, net, tree, and irregular. Two topologies, which do not use a radio band, have been developed for broadcast: ring, where the ends of the channel are connected, and bus, where they are not.

Voice traffic characteristics

To understand voice communications, we have to explore the history of voice networks. When AT&T was divested of its local phone companies on January 1, 1984, two distinct types of telephone companies were legally defined.

1. Local Exchange Carriers (LECs). These consist of the 23 Bell Operating Companies (BOCs) that were created by the divestiture, the former independent telephone companies such as GTE and Contel, and about 1500 small-town telephone companies.
2. Interexchange Carriers (IECs or sometimes IXCs), more often called simply long-distance carriers. These are the former AT&T Long Lines organization and other carriers, such as MCI and Sprint.

The United States was partitioned into Local Access and Transport Areas (LATAs). LATA boundaries conform more or less to the standard metropolitan statistical areas defined by the U.S. Department of Commerce. Originally, they conformed to the boundaries of the areas served by the BOCs.

LECs are prohibited from carrying inter-LATA calls and IECs are prohibited from carrying intra-LATA calls. Even when an LEC covers two LATAs, it must route the call through an IEC. An IEC is represented in a LATA by a Point of Presence (POP). A subscriber can connect to the IEC's facilities only at a POP. Typically, this connection is made by a line provided by the LEC from the subscriber's premises to the POP (A in the IEC diagram). Alternately, the subscriber can bypass the LEC and connect directly to the POP. For example, the subscriber could install a microwave link to the POP (B).

The providers of inter-LATA services fall into two classes, as follows:

Long-Distance Carriers. Companies that provide long-distance private lines or virtual private networks (switched lines that the subscriber uses in the same manner as dedicated private lines).

Packet Carriers. Companies that provide packet switching services. These companies often use the facilities of the long-distance carriers to construct their networks, of course.

Long-distance carriers compete with offerings in the following specific categories:

1. Voice Grade Service provides leased private analog lines.
2. Digital Data Service (DDS) provides leased digital lines used for data only at speeds ranging from 2400 bps to 19.2 kbps or, in some cases, 64 kbps. The DS0 Service provides 64-kbps digital lines, each providing one DS0 channel. Some carriers provide a clear channel capability, meaning that the carrier does not require any portion of the channel for control signaling. The channel is "clear," and the entire 64 kbps band width can be used by the subscriber. Otherwise, only 56 kbps of the 64 kbps can be used.
3. Fractional T1 lines are the FT1 lines described earlier in this chapter. Some vendors allow the individual FT1 components to be accessed as if they were private voice grade lines or DDS lines while others provide only T1 access.
4. T1 Service provides lines that support voice, data, or video at DS1 speed. Service may be channelized or unchannelized. Channelized means that it is divided into channels by the carrier who then takes care of multiplexing multiple DS0 channels onto it. If unchannelized, the subscriber is responsible for multiplexing.
5. T3 Service is like T1 service but at DS3 speed. It is generally unchannelized, though some vendors do offer channelized service.
6. Some carriers to international locations through certain gateway cities provide international private lines of various types. The subscriber must lease two lines, one to the gateway and one from the gateway to the international location.
7. Switched Data Services are often referred to as software-defined networks or virtual private networks (VPNs). They allow the subscriber to make use of the circuit switching capabilities of the carrier's long-distance facilities to control and monitor their private network. Typically, the carrier provides the subscriber with terminals that interface with the carrier's network control facility. The subscriber can monitor and reconfigure the VPN within certain limits established by the carrier.

With the advent of direct distance dialing (area codes) in the 1960s, it became possible to obtain a high-quality connection to any of over 100 million telephones in the United States by simply dialing the address (phone

number) of that device. This was the culmination of 100 years of development of electromechanical devices for the routing of calls and of analog devices for the transmission of voices. Voices were transmitted as a continuously varying electrical signal across a pair of wires (a local loop or a subscriber line loop) between the handset of the subscriber (the phone user) and an exchange or end office. (Before digital dialing, end offices had names such as Prospect or Elgin, so telephone numbers began with an abbreviation of the end office name, for example, PR6-6178 or EL3-1978.) An enterprise could install a Private Branch Exchange (PBX) to switch calls within the enterprise. PBXs were connected to local offices, which connected to trunks.

The electrical signal varied in frequency from about 600 to 3400 hertz (Hz, cycles per second). Although humans can hear frequencies from about 20 to 20,000 Hz, most speech energy is concentrated in the range of 600–3400 Hz.

Early telephone end offices accomplished switching with human operators, but in the 1930s switching was automated. Before digital dialing, electromechanical rotary switches (in smaller, especially rural, end offices) and crossbar switches connected subscriber lines. Telephones had rotary dials that generated electrical pulses, as many pulses as the number that was dialed (one pulse for "1," two pulses for "2," and so on through ten for "0"). The pulses "stepped" the switches to make the interconnection. Trunk lines connected end offices. As the number of end offices grew, it quickly became necessary to organize the phone system into a hierarchy; too many trunk lines would have been required to interconnect all of the end offices in even a metropolitan area, let alone the whole country (about 20,000 end offices are in use today). By connecting each end office to a toll center and connecting toll centers together with trunk lines, far fewer trunks were required even though any subscriber could reach any other subscriber in the area. Over time, a four-layer hierarchy was developed to interconnect all end offices in the United States.

Long-distance trunk lines were expensive to build, yet even the simple copper wires used for the first long-distance lines were capable of carrying an electrical signal with a much wider bandwidth than was required for voice transmission. A technique called FDM was developed that allowed a trunk line with only two pairs of wires to carry many voice conversations (many voice channels) simultaneously. Because of limitations in the technology first developed for FDM, a voice channel was standardized to be 4000 Hz of bandwidth. The additional 600 Hz between 3400 Hz and 4000 Hz was necessary to separate the channels on a multiplexed line so that one channel would not interfere with the next.

The addition of data

Today's digital telecommunications networks evolved out of the analog switched long-distance networks that were created in the early part of the 20th century. The telephone system was designed for voice communications (see Figure 3.1-7 Leased Lines diagram). Some errors can be introduced into a voice signal without causing problems for the people at either end of the conversation. For example, a shift in the phase of the signal (essentially a delay in the transmission of the signal) will have little or no effect on voice quality. But such errors can cause problems for data transmission. When the telephone network was an analog network, electromechanical switches could inject a certain amount of noise into circuits. Other instruments in the transmission path, such as multiplexers, could corrupt the signal further. At lower data rates, say 300 or 1200 bps, modems can correctly transmit data with few, if any, errors even when the line is noisy, but as data rates increase, noise causes more problems. As a practical

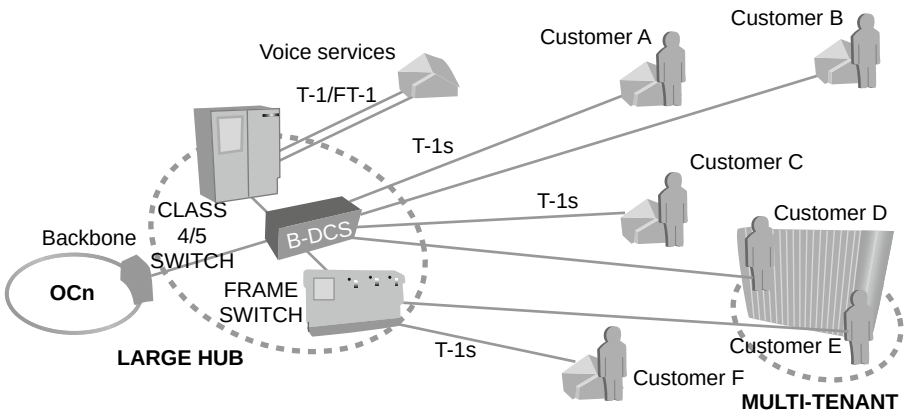


Figure 3.1-7 Voice communications network.

matter, in the analog network, data could not be transmitted reliably over long distances on switched lines at rates over 4800 bps. Therefore, it became common practice for the telephone companies to "lease" lines to companies for continuous, unswitched use. A company with two computer sites could, for example, lease a line or a set of lines to interconnect the sites.

Leased lines could be "conditioned" by the telephone company to make them usable at higher rates of transmission up to 19.2 bps. That basic limit was imposed by the previously mentioned 4000-Hz bandwidth of voice channels, although recent advances in error correction and data compression techniques have made effective rates significantly beyond that possible. A leased line would usually be composed, not of an actual dedicated set of copper wires, but of a dedicated FDM channel between area offices of the phone company. To obtain effective data transfer rates greater than 19.2 bps, it was then necessary for the subscriber to lease multiple lines and transmit data in parallel.

Leased lines have an additional characteristic that makes them desirable for certain uses: The length of the circuit is known (the telephone company usually is able to provide this information). The length does not usually vary from day to day. Fixed length is important to certain applications that have timing considerations related to actual circuit length, for example, channel extension. (Backup circuits can be leased for an additional charge. The length can vary when the backup is used.)

The data communication terms "half-duplex" and "full-duplex" are closely related to, but not synonymous with, the telephony terms "two-wire" and "four-wire. " A half-duplex circuit is one that provides transmission in two directions, but only in one direction at a time. A full-duplex circuit is one that provides transmission in both directions simultaneously. Obviously, and end-to-end four-wire circuit provides full-duplex capabilities. However, two-wire circuits can also be used for full-duplex communication by partitioning the available bandwidth into separate frequency bands for each direction of transmission (derived four-wire). This technique is often utilized when a full-duplex data communications circuit is desired over dial-up (two-wire) facilities. On the other hand, the existence of four-wire circuits does not necessarily imply that full-duplex transmission can be achieved through simultaneous use of both pairs. Very long-distance circuits require echo suppressers that effectively disable one pair of a four-wire circuit while the other pair is in use. Thus only one pair can be used at a time. Long-distance (four-wire) leased lines for data communications typically have the echo suppressers removed so that both pairs can be used simultaneously.

TDM architecture

TDM replaced FDM. TDM is a scheme in which numerous signals are combined for transmission on a single communications line or channel. Each signal is broken up into many segments, each having a very short duration. TDM multiplexers interleave the output of 12 codecs into a multiplex frame. Three standards prevail for multiplexing—North American, European, and Japanese. All three are based on a DS0 channel that is the pulse code modulation (PCM) output of a codec, and define five or six levels of successive multiplexing. They differ in the number of voice channels multiplexed onto a given level.

The circuit that combines signals at the source (transmitting) end of a communications link is known as a multiplexer. It accepts the input from each individual end user, breaks each signal into segments, and assigns the segments to the composite signal in a rotating, repeating sequence. The composite signal thus contains data from all the end users. At the other end of the long-distance cable, the individual signals are separated out by means of a circuit called a demultiplexer, and routed to the proper end users. A two-way communications circuit requires a multiplexer/demultiplexer at each end of the long-distance, high-bandwidth cable. If many signals must be sent along a single long-distance line, careful engineering is required to ensure that the system will perform properly.

This type of multiplexing combines many digital bit streams with relatively low bit frequency into a single bit stream with a relatively high bit frequency. It is, in essence, a way for many slow communications channels to "time share" a very fast channel. The advantage, of course, is that the cost per bit transmitted on a single fast channel is lower than on slower channels. In TDM, stations are allocated the entire bandwidth of frequencies but only for a small percentage of time (see Figure 3.1-8). TDM is accomplished by simply interleaving data from several bit streams. This can be done on a bit basis or on a byte basis (called bit interleaving and byte interleaving). During time interval 1, 8 bits from source channel 1 are transmitted. During successive intervals, bytes from successive source channels are transmitted on the output channel. A complete set of values from each input channel is called a frame.

Convergence technologies

Telecommunications convergence is the merger of legacy-based TDM architecture with today's packet-switching technology and call-control intelligence, which allows commercial carriers and service providers to consolidate voice and data networks to provide

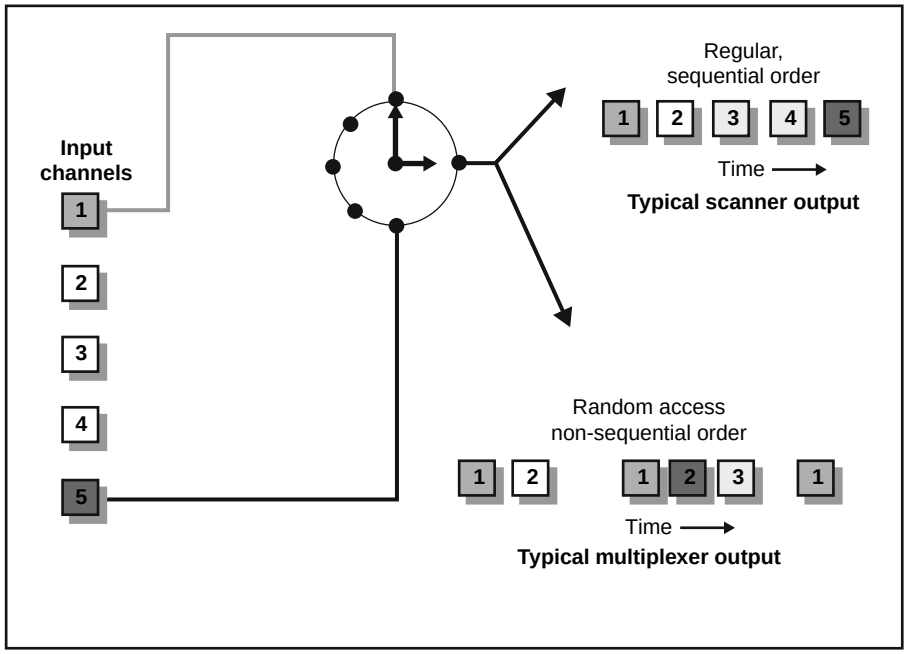


Figure 3.1-8 Time division multiplexing techniques.

integrated communications services. Convergence technologies are changing the way telecommunications companies will provide voice and data traffic. The public switched telephone network (PSTN) is one of the most reliable communications networks in existence. Using traditional PSTN services as an access point to the Internet has significantly encouraged the growth of data over costly Class-5 facilities, resulting in the need to reengineer this traditional TDM architecture. Convergence technologies will provide a packet-based architecture that combines the speed and efficiency of broadband with the full-featured signaling system 7 (SS7) architecture to create a hybrid network in which carriers and service providers can choose route options based on cost, efficiency, and fault management.

However, convergence is not accomplished by acquiring a single box. Those who intend to compete in the converged market space face significant challenges in choosing the correct equipment and software configured to provide them with the capabilities required to compete in today's Internet economy. Investors, entrepreneurs, and existing vendors must appreciate the complexities of convergence or risk the delays and unnecessary expense of adopting the wrong strategy.

Just as the Internet has predicated this shift to convergence, it has also provided the most compelling means of dominating in the converged market space.

With data and voice networks converging, it is apparent that circuit-switched networks built for voice communications are being adversely affected by the growth in dial data access. Rather than create a new class of equipment to divert data traffic from the PSTN, it is more prudent to develop strategies that leverage existing features found in the voice network today. What makes a converged network work is the ability to control both networks from one platform simultaneously while being able to manage call-control and multiple protocols from different switch applications.

The SS7 network is the vital link between the voice and data networks. SS7 enables routing and control of both voice and data traffic across converged networks while continuing to make intelligent network (IN) services available to the traditional world of telephony. Additionally, SS7 networks connect calls faster and reduce overall operational costs, which stands out as a distinct competitive advantage in a business where every penny counts. For instance, suppose a customer wants to dial a client in another region of the country. According to today's terms, it usually means dialing a 10-digit number that routes out of band (SS7) to its destination. In a converged network, the call would route into a host facility via SS7 A-links, where the host would first check for validation and provide a branded prompt recording, receive the listed directory number (LDN), and route the call via least-cost routing methods to its destination. Based on costs, the call could route out a packet-switched gateway into an ATM cloud to the distance end, where another gateway would receive the call and send it via PSTN to its destination. This routing architecture cuts out the cost of the transcontinental connection.

SS7 has long been a requirement for mainstream carriers but has been financially out of reach for many of the upstarts and growth carriers as a result of the high cost of deployment and licensing. Traditionally, SS7 has been managed by separate SS7 connections at each point of presence (PoP), but this is an expensive solution because of circuits and license fees from each location. A distributed network architecture, a solution that involves running a quad 4-system configuration, offers a cost-saving solution. SS7 servers connect to two disparate signal transfer points (STPs) in different cities for continuous uptime. The SS7 signaling is pulled into the host and distributed using transmission control protocol (TCP)/IP over a clear channel T1 WAN solution. The key advantage of this distributed SS7 solution lies in the ability to manage multiple sites from a single location in real time. This eliminates the need for multiple licenses for SS7 products and fees for each switch connection in the network.

There are many different applications that handle access to the SS7 network. There are also several issues to consider in choosing a solution. A key preference is for an open standards-based architecture that includes features such as platform compatibility with third-party database applications and a platform configuration tool (PCT) that can be used as a graphical interface to create and manage multiple configurations. The SS7 application should have the capability to be managed remotely through the same interfaces used to communicate with the rest of the platform. SS7 and IP signaling should both be supported, thus enabling these services to run across convergent voice and data networks.

Chapter 3.2

3.2

Network infrastructure

Juanita Ellis, Charles Pursell and Joy Rahman

One of the most important phases of a converged network is to choose the right technology vendor or network integrator, one that can provide a seamless solution without forklifting your existing network. A network manager or administrator should consider an open converged platform architecture so that it is easier to deploy converged and next-generation solutions without compromise. Converged communication services drive the current model and pace of business, enabling a whole new way of transacting commerce and interacting with customers and employees. The emergence of global competition and the continued domination of the Internet have driven a need for much more sophisticated multifunctional communication services in order for businesses to reach customers. Carriers are responding to the challenge through the building of much more versatile platforms. This occurs in both new network deployments and the transformation of their traditional singular communication networks such as asynchronous transfer mode (ATM), time division multiplexing (TDM), Frame Relay, from Internet Protocol (IP) to Multiservice Network Architectures.

Converged network supports all of the user's traffic types with packet-based protocols such as ATM, Frame Relay, or IP. Of these, the dominant protocol in the access, edge, and core of the network is the IP. The concept of a managed IP infrastructure for voice, video, and data transmission has begun to revolutionize the industry. It leverages the Internet so that enhanced services can be offered at lower costs and it is motivating the development of a family of dynamic, next-generation, real-time applications. For traditional carriers, the importance of the co-existence of existing and newer technologies and equipment is an important reality. Next-generation multiservice networks, based on packetized technologies, will not entirely push out existent circuit-switching equipment and solutions for many years. This is where a network element called a "media gateway" is needed—it performs both signaling and media format conversion between different network encoding and transmission standards. For new communications service providers that don't have a legacy of traditional networks, co-existence is moot. Flexible and versatile multiservice network architectures can provide an immediate competitive advantage. Although these are relatively new technologies, there are still issues of proven capability to be achieved. However, these new service providers must still work within existing networks to maximize customer coverage and appeal. The key to success in this new communications market is the ability to rapidly create new communications solutions that offer carriers and other communication services providers a competitive edge in attracting new customers as well as retaining existing ones. The adoption of open standards in equipment development generates a virtuous circle where a market for outsourced components fosters competition, which in turn drives innovation, so visionary network equipment makers are looking to feature and function development partners who adhere to open standards for leading-edge technology development.

Before discussing TDM, ATM, Frame Relay, IP and how a company can migrate voice over IP (VoIP), we need to understand the basic rules of traditional voice communications.

Voice networking

Basic voice technology has been available for more than 100 years. During that time, the technology has matured to the point at which it has become ubiquitous and largely invisible to most users. This legacy of slow

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2003 Elsevier Inc. All rights reserved

evolution continues to affect today's advanced voice networks in many ways, so it is important to understand the fundamentals of traditional voice technology before emulating it on data networks.

Traditional analog telephone instruments used for plain old telephone service (POTS) use a simple two-wire interface to the network. They rely on an internal two-wire/four-wire hybrid circuit to both transmit and receive signals. This economical approach has been effective but requires special engineering regarding echo.

Basic telephony

Three types of signaling are required for traditional telephony: supervision, alerting, and addressing. Supervision monitors the state of the instrument—for example, allowing the central office or PBX to know when the receiver has been picked up to make a call or when a call is terminated. Alerting concerns the notification of a user that a call is present (ringing) or simple call progress tones during a call (such as busy, ringback, and so on). Finally, addressing enables the user to dial a specific extension.

In addition to signaling, telephony services also provide secure media transport for the voice itself, analog-to-digital conversion, bonding and grounding for safety, power, and a variety of other functions when needed. Analog voice interfaces have evolved over the years to provide for these basic functions while addressing specific applications. Because basic POTS two-wire analog interfaces operate in a master/slave model, data equipment must emulate two basic types of analog interfaces: the user side and the network side. The user side (telephone) expects to receive power from the network as well as supervision.

A foreign exchange service (FXS) interface is used to connect an analog telephone, fax machine, modem, or any other device that would be connected to a phone line. It outputs 48 vdc power, ringing, and so on, and it accepts dialed digits. The opposite of an FXS interface is a foreign exchange office (FXO) interface. It is used to connect to a switching system providing services and supervision and, at the other end, it expects the switch to provide supervision and other elements. (Why "foreign"? The terms FXS and FXO were originally used within telephone company networks to describe provision of telephone service from a central office other than normally assigned.)

Within FXS and FXO interfaces, it is also necessary to emulate variants in supervision. Typical telephones operate in a loop start mode. The telephone normally presents a high impedance between the two wires. When the receiver goes off-hook, a low-impedance closed circuit is created between the two wires. The switch, sensing current flow, then knows that the receiver is off-hook and applies a dial tone. The switch also checks to be sure that the receiver is on-hook before sending a ringing signal. This system works well for simple telephones, but it can cause problems on trunks between PBXs and COs with high activity. In that situation, the remote end and the CO switch can both try to seize the line at the same time. This situation, called *glare*, can freeze the trunk until one side releases it. The solution is to short tip or ring to ground as a signal for line seizure, rather than looping it. This is called *ground start*.

After the line is seized, it is necessary to dial the number. Normal human fingers cannot outrun the dial receivers in a modern switch, but digits dialed by a PBX can. In that case, many analog trunks use a delay start or wink start method to notify the calling device when the switch is ready to accept digits.

Another analog interface often used for trunking is E&M. This is a four- or six-wire interface that includes separate wires for supervision in addition to the voice pair. *E&M* stands for "ear and mouth" or "Earth and magneto" and is derived from the early telephony days. The E&M leads are used to signal on-hook and off-hook states.

An analog interface works well for basic trunk connections between switches or PBXs, but it is uneconomical when the number of connections exceeds six to eight circuits. At that point, it is usually more efficient to use digital trunks. In North America, the T1 (1.544 Mbps) trunk speed is used, consisting of 24 digitized analog voice conversations. In other parts of the world, E1 (2.048 Mbps) is used to carry 30 voice channels. (Engineers refer to the adoption of E1 and T1 internationally as "the baseball rule"—there is a strong correlation of countries that play baseball to the use of T1. Therefore, the United States, Canada, and Japan have the largest T1 networks, while other countries use E1.)

The first step in conversion to digital is sampling. The Nyquist theorem states that the sampling frequency should be twice the rate of the highest desired frequency. Early telephony engineers decided that a range of 4000 Hz would be sufficient to capture human voices (which matches the performance of long analog loops). Therefore, voice channels are sampled at a rate of 8000 times per second, or once every 125 ms. Each one of these samples consists of an 8-bit measurement for a total of 64,000 bits per second to be transmitted. As a final step, companding is used to provide greater accuracy of low-amplitude components. In North America, this is μ-law (mu-law), while elsewhere it is typically A-law. For international interworking purposes, it is agreed that the North American side will make the conversion.

To construct a T1, 24 channels are assembled for a total of 1.536 Mbps, and an additional 8 bits are

added every 125 ms for framing, resulting in a rate of 1.544 Mbps. Often, T1 frames are combined into larger structures called SuperFrames (12 frames) and Extended-SuperFrames (24 frames). "Robbing bits" from the interior frames can then transmit additional signaling.

Basic T1 and E1 interfaces emulate a collection of analog voice trunks and use robbed bit signaling to transfer supervisory information similar to the E&M analog model. As such, each channel carries its own signaling, and the interface is called channel-associated signaling (CAS). A more efficient method uses a common signaling channel for all the voice channels. Primary Rate Interface for ISDN is the most typical example of this common channel signaling (CCS). If voice/data integration is to be successful, all of these voice interfaces must be supported to provide the widest possible range of applications. Over the years, users have grown to expect a certain level of performance, reliability, and behavior of a telecommunications system, which must be supported while going forward. All these issues have been solved by various packet voice systems today so that users can enjoy the same level of support to which they have become accustomed. Before we jump into ATM, Frame Relay, and IP, we need to step back and take a look at TDM.

Time division multiplexing

TDM is a scheme in which numerous signals are combined for transmission on a single communications line or channel. Each signal is broken up into many segments, each having a very short duration.

The circuit that combines signals at the source (transmitting) end of a communications link is known as a multiplexer. It accepts the input from each individual end user, breaks each signal into segments, and assigns the segments to the composite signal in a rotating, repeating sequence. The composite signal thus contains data from all the end users. At the other end of the long-distance cable, the individual signals are separated out by means of a circuit called a demultiplexer and then routed to the proper end users. A two-way communications circuit requires a multiplexer/demultiplexer at each end of the long-distance, high-bandwidth cable.

If many signals must be sent along a single long-distance line, careful engineering is required to ensure that the system will perform properly. An asset of TDM is its flexibility. The scheme allows for variation in the number of signals being sent along the line, and constantly adjusts the time intervals to make optimum use of the available bandwidth. The Internet is a classic example of a communications network in which the volume of traffic can change drastically from hour to hour. In some systems, a different scheme, known as frequency division multiplexing (FDM), is preferred.

Voice over ATM

Many companies have ATM backbone infrastructure and deploying convergence within their network and outside their network. It is crucial to understand how an ATM network is configured before even considering convergence for voice/video/data. ATM is a dedicated-connection switching technology that organizes digital data into 53-byte cell units and transmits them over a physical medium using digital signal technology. Individually, a cell is processed asynchronously relative to other related cells and is queued before being multiplexed over the transmission path. Because ATM is designed to be easily implemented by hardware (rather than software), faster processing and switch speeds are possible. The pre-specified bit rates are either 155.520 Mbps or 622.080 Mbps. Speeds on ATM networks can reach 10 Gbps. Along with Synchronous Optical Network (SONET) and several other technologies, ATM is a key component of broadband ISDN (BISDN).

From the start, ATM was designed to be a multimedia, multiservice technology. It has been accepted by the marketplace for its ability to deliver high-speed data services. Until the recent past, its potential for deploying for voice services was overlooked. With the competitiveness of today's market, the network operators and the service providers have been continuously striving to reduce operating costs and lift network efficiency. They recognized that significant economic benefits could be achieved once the data traffic and voice traffic were integrated onto a single network. Because ATM has been around for a decade or more, claiming to be a multimedia technology, most of the service providers have started installing single ATM infrastructure to support voice, video, and data transfer. Initially, there were a lot of technical issues that were unaddressed, which basically hampered the growth of VTOA. Thanks to the efforts of the ATM Forum and its members, these issues have been addressed and it is now possible to build and operate an ATM network to meet the needs of various types of voice application.

Many in the telecommunications industry believe ATM will revolutionize the way networks are designed and managed because ATM combines the best features of two common transmission methods. Its connection-oriented nature makes ATM a reliable service for delay-sensitive applications such as voice, video, or multimedia. Its pliant and efficient packet switching provides flexible transfer of other forms of data. In a relatively short period of time, ATM has gained

a worldwide reputation as the ultimate means of solving end-to-end networking problems. The popularity of ATM has grown such that virtually every LAN equipment vendor and service provider is racing to develop ATM-based products.

Voice over ATM (VoATM) can be supported as standard pulse code modulated (PCM) voice via circuit emulation (AAL1, described later) or as VBR voice in ATM cells as AAL2 (also described later). ATM offers many advantages for transport and switching of voice. First, quality of service (QoS) guarantees can be specified by service provisioning or on a per-call basis. In addition, call setup signaling for ATM switched virtual circuits (SVCs), Q.2931, is based on call setup signaling for voice ISDN. Administration is similar to circuit-based voice networks.

However, VoATM suffers from the burden of additional complexity, and incomplete support and interoperability among vendors. It also tends to be more expensive because it is oriented toward all-optical networks. Most importantly, ATM is typically deployed as a WAN Layer 2 protocol and therefore does not extend all the way to the desktop. Nevertheless, ATM is quite effective for providing trunking and tandem switching services between existing voice switches and PBXs.

Voice over Frame Relay (VoFR) has become widely deployed across many networks. Like VoATM, it is typically employed as a tie trunk or tandem-switching function between remote PBXs. It benefits from much simpler administration and relatively lower costs than VoATM, especially when deployed over a private WAN network. It also scales more economically than VoATM, supporting links from T1 down to 56 kbps. When deployed over a carefully engineered Frame Relay network, VoFR works very well and provides good quality. However, voice quality over Frame Relay can suffer depending on network latency and jitter. Although minimal bandwidth and burstiness are routinely contracted, latency and jitter are often not included in service level agreements (SLAs) with service providers. As a result, voice performance can vary. Even if quality is good at first, voice quality can degrade over time as a service provider's network becomes saturated with more traffic. For this reason, many large enterprise customers are beginning to specify latency and jitter, as well as overall packet throughput from carriers. In these situations, VoFR can provide excellent service.

VoIP has begun to be deployed in recent years as well. Unlike VoFR and VoATM, VoIP is a Layer 3 solution. It offers much more value and utility because IP goes all the way to the desktop. This means that in addition to providing basic tie trunk and tandem-switching functions to PBXs, VoIP can actually begin to replace those PBXs as an application. As a Layer 3 solution, VoIP is routable and can be carried transparently over any type of network infrastructure. This includes both Frame Relay and ATM. Of all the packet voice technologies, VoIP has perhaps the most difficult time supporting voice quality because QoS cannot be guaranteed. Normal applications such as TCP running on IP are insensitive to latency but must retransmit lost packets due to collisions or congestion. Voice is much more sensitive to packet delay than packet loss. In addition to normal traffic congestion, QoS for VoIP is often dependent on lower layers that are ignorant to the voice traffic mingled with the data traffic.

The ATM Forum and the ITU have specified different classes of services to represent different possible traffic types for VoATM. Designed primarily for voice communications, constant bit rate (CBR) and VBR classes have provisions for passing real-time traffic and are suitable for guaranteeing a certain level of service. CBR, in particular, allows the amount of bandwidth, end-to-end delay and delay variation to be specified during the call setup. Designed principally for bursty traffic, unspecified bit rate (UBR) and available bit rate (ABR) are more suitable for data applications. UBR, in particular, makes no guarantees about the delivery of the data traffic.

The method of transporting voice channels through an ATM network depends on the nature of the traffic. Different ATM adaptation types have been developed for these different traffic types, each with its own benefits and detriments. ATM adaptation layer 1 (AAL1) is the most common adaptation layer used with CBR services.

Unstructured AAL1 takes a continuous bit stream and places it within ATM cells. This is a common method of supporting a full E1 byte stream from end to end. The problem with this approach is that a full E1 may be sent, regardless of the actual number of voice channels in use. (An EI is a wide-area digital transmission scheme used predominantly in Europe, which carries data at a rate of 2.048 Mbps.)

Structured AAL1 contains a pointer in the payload that allows the digital signal level 0 (DS0) structure to be maintained in subsequent cells. This allows network efficiencies to be gained by not using bandwidth for unused DS0s. (A DS0 is a framing specification used in transmitting digital signals over a single channel at 64 kbps on a T1 facility.)

The remapping option allows the ATM network to terminate structured AAL1 cells and remap DS0s to the proper destinations. This eliminates the need for permanent virtual circuits (PVCs) between every possible source/destination combination. The major difference from the previous approach is that a PVC is not built across the network from edge to edge.

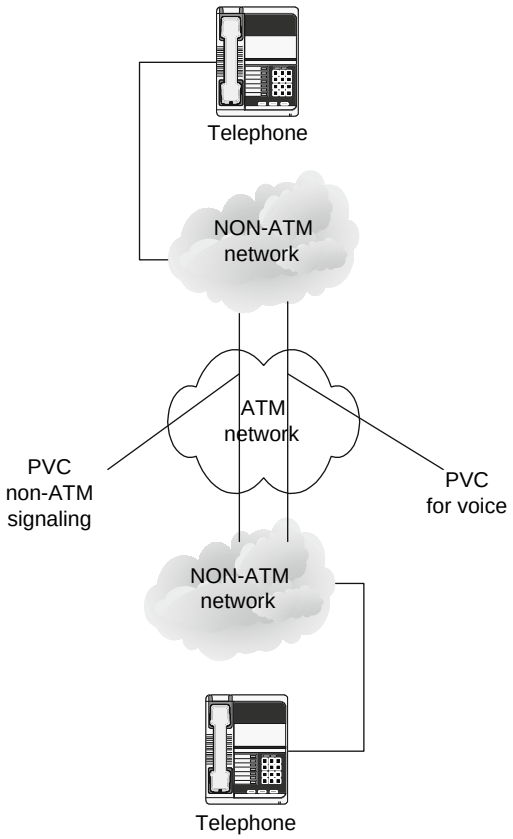


Figure 3.2-1 The VoATM signaling transport model describes the transport method, in which voice signaling is carried through the network transparently.

VoATM signaling

Figure 3.2-1 describes the transport method in which voice signaling is carried through the network transparently. PVCs are created for both signaling and voice transport. First, a signaling message is carried transparently over the signaling PVC from end station to end station. Second, coordination between the end systems allows the selection of a PVC to carry the voice communication between end stations.

At no time is the ATM network participating in the interpretation of the signaling that takes place between end stations. However, as a value-added feature, some products are capable of understanding CAS and can prevent the sending of empty voice cells when the end stations are on-hook.

VoATM addressing

ATM standards support private and public addressing schemes. Both schemes involve addresses that are 20 bytes in length (shown in Figure 3.2-2).

The *Authority and Format Identifier* (*AFI*) identifies the particular addressing format employed. Three identifiers are currently specified: data country code (DCC), international code designator (ICD), and E.164. A standard body administers each one. The second part of the address is the initial domain identifier (IDI). This address uniquely identifies the customer's network. The E.164 scheme has a longer IDI that corresponds to the 15-digit ISDN network number. The final portion, the domain-specific part (DSP), identifies logical groupings and ATM end stations.

In a transport model, you don't need to be aware of the underlying addressing used by the voice network. However, in the translate model, the capability to communicate from a non-ATM network device to an ATM network device implies a level of address mapping. Fortunately, ATM supports the E.164 addressing scheme that is employed by telephone networks throughout the world.

VoATM routing

ATM uses a *private network-to-network interface* (*PNNI*), a hierarchical link-state routing protocol that is scalable for global usage. In addition to determining reachability and routing within an ATM network, it is also capable of call setup.

A virtual circuit (VC) call request causes a connection with certain QoS requirements to be requested through

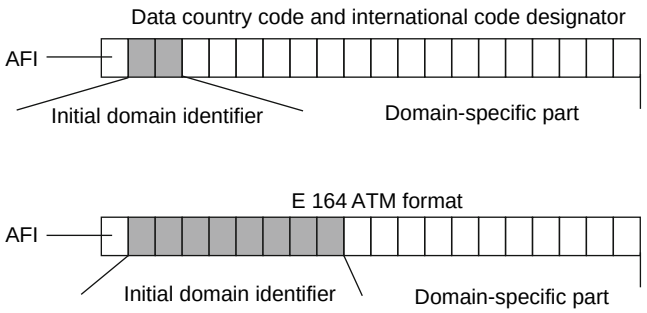


Figure 3.2-2 ATM supports a 20-byte addressing format.

the ATM network. The route through the network is determined by the source ATM switch. This is based on what it determines is the best path through the network and on the PNNI protocol and the QoS request. Each switch along the path is checked to determine whether it has the appropriate resources for the connection.

When the connection is established, voice traffic flows between end stations as if a leased line existed between the two. This specification spells out routing in private networks. Within carrier networks, the switch-to-switch protocol is B-ICI. Current research and development of integrated non-ATM and ATM routing will yield new capabilities to build translate-level voice and ATM networks.

VoATM and delay

ATM has several mechanisms for controlling delay and delay variation. The QoS capabilities of ATM allow the specific request of CBR traffic with bandwidth and delay variation guarantees. The use of VC queues allows each traffic stream to be treated uniquely. Priority can be given for the transmission of voice traffic. The use of small, fixed-size cells reduces queuing delay and the delay variation associated with variable-sized packets.

Benefit of VoATM

With the standards now in place, it is possible for packet switching techniques such as Frame Relay or ATM to deliver high-quality speech. Some of the intrinsic advantages ATM has over other network technologies are listed here.

- The very format of the cell was arrived at by considering data, voice, and video payload requirements. ATM cells are of fixed size, 53 bytes each with 48 bytes for payload and 5 bytes for ATM header. This helps in reducing the packetization delay significantly, which is one of the major delay parameters.
- It supports extensive QoS, which allows voice traffic to be transmitted across the network in a reliable jitter-free way.
- Various AALs support various service classes capabilities.
- ATM switches have always been designed with effective traffic management capabilities—for example, call admission control, usage parameter control, traffic shaping, etc.
- It allows a single network for voice, data, and video.
- Interworking with PSTN is relatively straightforward.

Though ATM is equipped with transferring voice over the network efficiently, ATM onto the desktop would not be popular enough until the turn of the millennium. The reasons being, there are enough competing technologies, like 100-Mbps Ethernet/Gigabit Ethernet that provide similar services with minimal infrastructure upgrade to be done. But, when WAN is considered, ATM has its own niche over its competing technologies for the same reasons previously discussed. While designing and engineering voice over an ATM WAN, there are sets of design issues that need to be addressed. Some of them are as follows.

Technical challenges

A packetized approach to transmit voice faces a number of technical challenges, which spring from the real-time or interactive nature of the voice traffic. Some of the challenges that need to be addressed are discussed next.

Echo

This is a phenomenon where the transmitted voice signal gets reflected back. It is due to the unavoidable impedance mismatch and the four-wire/two-wire conversion between the telephone handset and the communication network. It can, depending on the severity, disrupt the normal flow of conversation. Its severity depends on the round-trip time delay. It is found that if the round-trip time delay is more than 30 ms, the echo becomes significant, making it difficult to have a normal conversation.

End-to-end delay

Voice is most sensitive to delay and mildly sensitive to variations in delay (jitter). It is highly critical that the delay is kept at a bare minimum to hold an interactive communication end-to-end. It has been found that delay can have two effects on communication performance. Delay can interfere with the dynamics of voice communication, in the absence of noticeable echo, whereas in the presence of noticeable echo, increasing delay makes echo effects worse. When the delay reaches above 30 ms, echo cancellor circuits are required to control the echo. Once the echo cancellor circuits are in place, network delays can be allowed to reach up to 150 ms without further degrading the voice quality.

According to the ITU-T Recommendation G.114, Table 3.2-1 shows allowable delay limits for one-way

Table 3.2-1 Delay limits.

Delay	Acceptability
0–150 ms	Acceptable to most user applications
150–400ms	Acceptable when the impact on user's application is minimal
More than 400 ms	Unacceptable

transmission time for connections with adequate echo control.

Delay occurs in ATM networks because of one or more of the following reasons:

a) Packetization delay (or cell construction delay)
This is the time taken to fill in a complete packet/cell before it is transmitted. Normal PCM encoded voice samples arrive at the rate of 64 Kbps, which means it takes around 6 ms to fill the entire 48-byte payload of the ATM cell. The problem can be addressed either with partially filled cells or by mulitplexing several voice calls into a single ATM VCC (Virtual Circuit Channel).

b) Buffering delay
Sometimes, due to the delay in transit, some cells might arrive late. If this happens, SAR (Segmentation and Reassembling) function provided by the Adaptation layer might have to under-run with no voice data to process, which would result in gaps in the conversation. To prevent this the receiving SAR function would accumulate a buffer of information before starting the reconstruction. In order to ensure no under-runs occur, the buffer size should be kept in such a way it exceeds the maximum predicted delay. The size of the buffer translates into delay, as each cell must progress through the buffer on arrival at the emulated circuit's line rate. This implies that the Cell Delay Variation (CDV) has to be controlled within the ATM network.

c) Encoding delay
This is the processing time taken by the compression algorithms to encode the analog signal to digital form.

Silence suppression

Voice is inherently variable. It is found that on an average, human voice has a speech activity factor of about 42 percent. There are pauses between sentences and words with no speech in either direction. Also, voice communication is a half-duplex, that is, one person is silent while the other speaks. One can take advantage of these two characteristics to save bandwidth by halting the transmission of cells during these silent periods. This is known as silence suppression.

Compression algorithms

ADPCM (Adaptive Differential Pulse Code Modulation) and CELP (Code Excited Linear Prediction) are the two major compression algorithms that are used. Now, LD-CELP (Low Delay CELP), a derivative of CELP, is the compression algorithm most commonly used on any voice signal. This has been standardized by ITU as ITU G.728. This provides a toll quality voice at 16 kbps with low encoding/decoding delay. Table 3.2-2 compares the various compression techniques for their bandwidth usage, MIPS (millions of instructions per second), and delay.

Signaling

This relates to the efficient utilization of resources and the transfer of control and signaling information. There are two parts in a voice call—the actual voice samples and the signaling information, such as dialed number, the on-hook/off-hook status of the call, and other routing and control information. This signaling can be encoded and may be sent as CCS, where signaling information from different channels is aggregated into a single signaling channel, and CAS, where signaling information is embedded within each discrete voice channel.

Synchronization

The transport of voice demands that the data be synchronized between the speaker and the listener. There are two standard mechanisms that are used to achieve synchronization between point-to-point applications. They are Adaptive Clocking and Synchronous Residual Time Stamping (SRTS). These mechanisms work by

Table 3.2-2 Comparison between various compression algorithms.

Algorithm	Bandwidth	MIPS (C5x DSP)	Total Codec Delay (msec)	Application
PCM	64 kbit/s	n/a	0.25	PSTN
ADPCM (G.726)	32 kbit/s	10	0.25	PSTN, cordless phones
CS-ACELP (G.729)	8 kbit/s	30	25	VoFR, VoATM, VoIP
CS-ACELP Annex A (G.729A)	8 kbit/s	20	25	VoFR, VoATM, VoIP
LD-CELP (G.728)	16 kbit/s	40	1.25	PSTN
MP-MLQ (G. 723.1)	5.3/6.3 kbit/s	30	67.5	Multimedia and VoIP

adjusting the clock rate at one end of the circuit based on the clock rate of the other end.

The above-mentioned mechanisms work effectively only in the master–slave environment or point-to-point communication. When multipoint services are in operation, it is not possible for a slave to adjust its clock based on two or more difference signals coming from different master sites. For multipoint service, it is easy to adopt an externally synchronized model where each node in the network is synchronized to some external clock source.

Standards and specifications of ATM

Various applications are available for the transport of voice over an ATM network. Each application has differing requirements for voice transport based on what class of network operators they are defined in. National or International Operators typically have an extensive PSTN service operating over SDH/SONET or PDH infrastructure. When bandwidth is limited, there will be a requirement to integrate voice and data traffic, for reasons of efficiency, into a single ATM network. Within the local loop, ATM may be a valuable solution for the carriage of voice and data to business premises.

Alternate Carriers or Value-Added Network Suppliers take up licenses to provide communication services in competition with the incumbent national operators. Instead of having their own transmission infrastructure, they buy bandwidth from the primary operator. Cost and limited availability of bandwidth demand ATM efficiency as well as the integration of voice and data services. An example of alternate carriers is cellular phone operators.

Private/Enterprise networks buy bandwidth at commercial (retail) rates and achieve the most they can with the resources on hand. Such organizations will have already deployed a TDM network utilizing E1 or T1 links. They will be looking to integrate these solutions into a new ATM network and gain improvements in network performance and efficiency by moving from TDM to statistical multiplexing. With this scenario, two voice transport models have come up. One is known as "voice trunking" and the other as "voice switching."

a) Voice trunking

This involves tunneling of voice traffic over the ATM network between two fixed end points. This is an appropriate mechanism for connecting of voice switch sites, PBXs, or message switching centers. Here, network is not needed to process or terminate signaling, other than the opportunity to use the signaling to detect idle channels.

b) Voice switching

Here, the ATM network interprets the voice signaling information and routes the call across the network. In voice switching, the ATM switch receives a call and routes it to the appropriate destination. The VPN network fits appropriately for this kind of functionality. This type of network solution requires that ATM networks interpret the signaling provided from the voice network. Previously, this posed a major challenge as the signaling standards were proprietary. At the present, many vendors provide ATM-based solutions that are able to interpret the signaling provided by their own voice switches. Widespread adoption of ISDN and QSIG voice signaling standards are allowing ATM vendors to offer a standardized voice signaling interface.

From the foregoing analysis of various network operators and transport models, one could see a common set of network requirements emerge. These demonstrate that the minimum characteristics shown in Table 3.2-3 have to be supported to implement voice trunking.

Apart from these basic demands, to support a complex voice-switched solution or implement efficient statistical mulitplexing, the requirements shown in Table 3.2-4 also have to be met.

All the proposed ATM solutions will be measured against the current generation of TDM solutions that are successfully deployed. ATM Forum's VTOA workgroup, which is actively involved in developing standards for VoATM networks, has developed a set of solutions or standards that could offer direct commercial or operational benefits to any user. These standards allow voice traffic to be carried over an ATM network more efficiently than any packet- or TDM-based infrastructures.

The ATM Forum has defined three principal approaches to carrying voice over an ATM network. These are

- Circuit emulation services (CES), which are used to carry full or fractional E1/T1 circuits between end points.

Table 3.2-3 Objectives to be achieved—1.

Characteristics	Necessity
Adaptation	A mechanism to encode voice samples into ATM while meeting the delay and real-time constraints of voice traffic
Signaling	A mechanism that allows the end-to-end transport of voice signaling (Common Channel or Channel Associated) with the voice traffic
Low cross network delay (latency)	To minimize delay issues and allow normal interactive conversation (This is not a requirement for broadcast applications)
Limited variation in delay	To minimize delays and allow effective echo cancellation

Table 3.2-4 Objectives to be achieved—2.

Characteristics	Necessity
Signaling analysis	To allow set up and tear-down of circuits on demand (or allocation and release of resources)
Call switching and routing mechanisms	To allow configuration of "real-world" VPN applications
Silence suppression or VBR encoding	To realize statistical gain (provides at least a doubling in performance)
Call Admission Control (CAC)	To ensure QoS is preserved
Network resource utilization	To allow statistical overbooking of network resources

- Dynamic bandwidth CES (DB-CES).
- ATM trunking of narrowband services using AAL2.

Circuit Emulation Services

Circuit Emulation allows the user to establish an AAL1 connection to support a circuit, such as a full T1 or E1, over the ATM backbone. In using CES, the ATM network provides a transparent transport mechanism for various CBR services based on AAL1. It specifically covers the following types of CBR services:

1. Structured DS1/E1 n * 64 kbps (fractional DS1/E1) service.
2. Unstructured DS1/E1 (1.544 Mbps, 2048 Mbps) service.
3. Unstructured DS3/E3 (44,736 Mbps, 34,368 Mbps) service.
4. Structured J2 n * 64 kbps (fractional J2) service.
5. Unstructured J2 (6312 Mbps) service.

Figure 3.2-3 shows two ATM circuit emulation services (CES) interworking functions (IWFs) connected to an ATM network via physical UNI interfaces. The other sides of the CES-IWFs are connected to standard CBR circuits (e.g., DS1/DS3, J2, or E1/E3), which may originate, for example, on a user's PBX. The job of the two IWFs is to extend this CBR circuit over the ATM network. This means an ATM portion of the connection should retain the bit integrity— that is, analog signal loss cannot be inserted and voice echo control cannot be performed. These must be performed either by the DTE or before the ATM CES IWF is encountered. An ATM UNI physical interface has two characteristics that are relevant when supporting CES.

1. Bandwidth—The ATM interface must provide adequate bandwidth to carry n * 64 or unstructured traffic after segmentation.
2. Timing—The ATM interface can be used to convey timing traceable to a primary reference source from the ATM network to the CES inter working function, where the external connection to network timing is not supported.

An ATM network providing CES should also consider QoS parameters such as Peak Cell Rate (PCR) and CDV. There are standards that specify what values are optimum for these parameters.

CES's advantage is the simplicity of implementation. The ATM network is used to provide virtual replacements for physical links in an existing network. Still, CES has two limitations. First, it is unable to provide any statistical multiplexing. It does not differentiate between idle and active timeslots. This means all idle traffic is carried. Therefore, CES voice transport consumes 10 percent

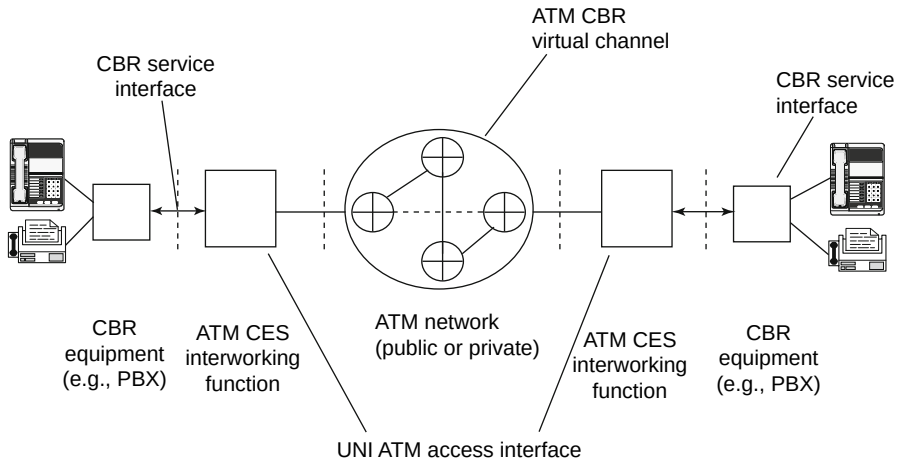


Figure 3.2-3 ATM circuit emulation.

more bandwidth than would be required to transfer the same voice traffic over leased circuits. Second, it is often implemented as a point-to-point service, providing the transport of the contents of one network physical interface to another physical network interface. This can prevent the implementation of some network topologies, and can result in increased network cost.

Dynamic Bandwidth CES

The limitations in CES resulted in the development of a new standard from the VTOA workgroup of the ATM Forum, which is referred to as "Dynamic Bandwidth Circuit Emulation Services DB-CES." The objective of this standard is to detect active or inactive timeslots of a legacy TDM trunk from a PBX or multiplexer so that inactive timeslots can be dropped from the next ATM structure to allow you to use and reutilize this bandwidth for other services such as CBR, VBR, UBR, and ABR applications.

ATM trunking of narrowband services using AAL2

The foregoing CES mechanisms treat voice as being a constant stream of information encoded as a CBR stream. But in the real sense, voice is a combination of talk spurts and silence. So bandwidth is wasted when transmitting the silence. These mechanisms also typically minimize the problems of cell construction delay by transmitting the voice as an uncompressed 64 kbps. Such approaches deny the network operator a chance to gain significant bandwidth from voice compression technologies. To address these limitations, the ATM Forum came up with the advanced mechanism for the transport of voice as a VBR compressed stream.

AAL2 is a new ATM Adaptation Layer, specified in ITU-T Recommendation I.363.2 (1997) with the specific mandate to provide efficient VoATM services. AAL2 supports, in addition to those of AAL1, the following features:

- Efficient bandwidth usage through VBR traffic.
- ATM bandwidth reduction support for voice compression, silence detection/suppression, idle voice channel deletion.
- Multiple voice channels with varying bandwidth on a single ATM connection.

Thus AAL2 can enable voice applications by using higher-level layer requirements such as voice compression, silence detection/suppression, and idle channel removal. AAL2 is divided into two sub-layers: the Common Part Sub-layer (CPS) and the Service Specific Convergence Sub-layer (SSCS).

a) AAL2 Common Part Sub-layer

CPS provides the basic structure for identifying the users of the AAL, assembling/disassembling the variable payload associated with each individual user, error correction, and the relationship with the SSCS. Each AAL2 user can select a given AAL-SAP associated with the QoS required to transport that individual higher-layer application. AAL2 makes use of the service provided by the underlying ATM layer. Multiple AAL connections can be associated with a single ATM layer connection, allowing multiplexing at the AAL layer.

b) AAL2 Service-Specific Convergence Sub-layer

According to the recommendations of the ITU-T I.363.2, the SSCS is defined as the link between the AAL2 CPS and the higher-layer applications of the individual AAL2 users. Several SSCS definitions that take advantage of the AAL2 structure for various higher-layer applications are planned. A null SSCS, already understood and used in conjunction with the AAL2 CPS, satisfies most mobile voice applications. To satisfy higher-layer requirements associated with data and AAL2 configuration messages—called AAL2 Negotiation Procedures, or ANP—an SSCS for segmentation/reassembly (temporarily called I.SEG) is in development within the ITU-T Study Group 13.

Let's discuss the ATM Trunking standard in brief—an important concept to understand in order to have a successful VoATM network.

a) Switched trunking

Switched trunking involves analysis of the signaling that accompanies an incoming narrowband call and routing of its bearer information to an AAL2 channel within a VCC between IWFs. Once the narrowband call has ended, subsequent calls occupying the same narrowband channel (TDM timeslot) may be switched to different AAL2 channels and VCCs. In other words, there is no permanent relationship between a narrowband channel and an AAL2 channel.

b) Non-switched trunking

In non-switched trunking, the information stream of a narrowband channel is always carried on the same AAL2 channel within the same VCC. In other words, there is a permanent correspondence between a narrowband channel and the AAL2 channel and VCC designated for its support. Non-switched trunking involves no termination of signaling and no routing of narrowband calls in the IWFs.

The interworked voice and data network of the future promises the best of all network worlds: the installed base of Frame Relay, the speed and quality of ATM, and

the ubiquity of IP. Currently, fragmentation techniques in Frame Relay, IP, and ATM are quite similar, but prioritization techniques, signaling protocols, and voice compression algorithms are not compatible. Progress is being made toward standardization within each protocol and interworking between them, but considerable work still remains.

Because a comprehensive standard has not been adopted for any one technology, it is unrealistic to expect the emergence of interoperability standards between technologies in the near future. Interworking solutions will therefore have to be proprietary. It is essential that the interoperability be transparent to the users, who want to communicate throughout the network efficiently and without concern for the technological issues involved. Due to the lack of interoperability standards for voice communications over Frame Relay, IP, and ATM, the vendors must develop proprietary interworking solutions.

There are many situations in which interworking between technologies is required within a corporate network. For example, corporations that are running data and VoFR might require VoIP to extend the network to remote locations that don't have a Frame Relay infrastructure without deploying additional equipment. This may also be required for telecommuters working from home, salespeople working from hotel rooms, and resellers that want to access information.

Voice over Frame Relay

In the last few years, data networks have been growing at a much faster rate than voice networks, mainly due to the growth of the Internet. Soon the amount of data traffic will exceed that of voice traffic. As a result of this trend, more and more voice is being sent over data networks (VoFM, VoIP, and VoATM) rather than data being sent over voice networks (via V.34 and V.90 modems).

When Frame Relay was introduced in the early 1990s, the data technology was not originally designed to carry voice. Despite valid reservations about the reliability of voice over frames, the promise of "free voice" eventually proved too alluring. Soon users were experimenting with transporting voice over their Frame Relay devices while equipment vendors worked overtime to make the promise of quality VoFR a reality.

As the public Internet exploded in the mid-1990s and users began implementing IP-based networks, the call for VoIP grew louder. Here, too, equipment manufacturers are developing products to enable inexpensive, universal voice over data networks.

Carriers, however, were caught in a dilemma. Could they afford to cannibalize their highly profitable public switched telephone network? Could they not afford to capitalize on the demand for digital voice? The drama is just unfolding.

Although significant progress has been made in engineering packet networks (Frame Relay, IP, and ATM) to carry voice as well as data, today's market is demanding a true convergence of these technologies into a single and ubiquitous communications service without being limited by the underlying technology. The next challenge, then, is to develop interconnecting and interworking standards in order to deliver voice services ubiquitously over Frame Relay, IP, and ATM.

VoFR service is enabled by the use of Frame Relay communications devices, such as routers or FRADs, configured with voice modules. These devices are sometimes referred to as Voice FRADs, or VFRADs.

Although VoFR implementations are efficient and cost effective for intra-enterprise or intra-corporate communication, there are considerations, such as quality, to be made regarding the use of packetized voice when dealing with non-Frame Relay WAN, or off-net, users. Figure 3.2-4 illustrates how a typical VoFR implementation might look, supporting simultaneously both on-net (VoFR) and off-net (POTS) implementations.

Equipping your company's network for VoFR

You can deliver voice over any carrier's frame relay services, assuming that you have ensured that the carrier's frame relay network can guarantee a delay and jitter that adequately meets your company's needs. Most of the major carriers offer frame relay services, including MCI WorldCom and AT&T Solutions. Few carriers offer managed VoFR services. However, AT&T Solutions announced a managed VoFR service called *Managed Multiservice Networking* (*MMN*). Unless you use a managed VoFR service, such as that offered by AT&T Solutions, you will have to purchase your own VoFR equipment. Most of the major router vendors, including 3Com Corp., Avaya Inc., Lucent Technologies, Cisco Systems Inc., and Nortel Networks Corp., offer routers, switches, or voice-enabled frame relay access devices (VFRADs).

Voice switching is the technology that enables us to dial a telephone number and reach the desired destination. It implements the signaling required to establish, maintain, and tear down a voice call. Voice switching methods affect the speed, efficiency, and quality of voice calls traveling through the network. Therefore, before purchasing VFRADs, users must consider the voice switching capabilities of each product, including ease of configuration, flexibility in dialing plans, quality of compressed voice, integrated fax support, and support for

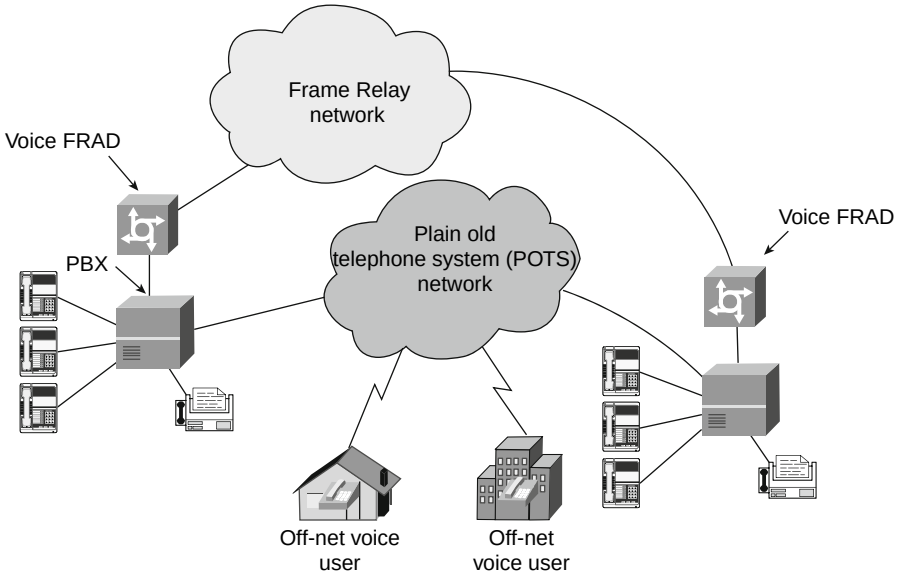


Figure 3.2-4 Example of VoFR.

various physical interfaces and signaling methods. Before explaining the importance of VoFR, it is necessary to describe the methods implemented to enable the Frame Relay network to carry voice as well as data. In particular, some of the common voice-enabling techniques incorporated into VFRADs are described in the following sections.

The nature of the data network and its implications for voice

Frame Relay is known as packet-switching technologies. This is in contrast to the public telephone network, which is a circuit-switching technology, designed to carry voice transmissions. Frame Relay inserts data into variable-sized frames or packets.

The packet-switching and cell-switching networks perform statistical multiplexing. That is, they dynamically allocate bandwidth to various links based on their transmission activity. Because bandwidth is not reserved for any specific path, the available bandwidth is allotted according to network needs at any particular time.

Compare this to the traditional voice (or circuit-switching) network, in which a path is dedicated to the transmission for the duration of the call, which is sent in a continuous bit stream. The line is monopolized by a call until it is terminated, even when the caller is put on hold and during periods of silence. Although this guarantees reliable and immediate transmission of voice, it results in very inefficient use of bandwidth. A line that is dedicated to the telephone cannot be utilized by other data even when there are no voice transmissions.

Originally designed to handle bursty data traffic, packet-switching networks (except for ATM) are inherently less efficient than the circuit-switching network in dealing with voice. To achieve good voice quality, the delay of voice packets across the network must be minimal and fixed. Due to the shared nature of the packet/cell-switching network, it might take time for transmissions to travel across the network. A transmission can be delayed because of network congestion. For example, it might "get stuck" behind a long data transmission that delays other packets. Network congestion can also result in dropped packets, which also detrimentally affects the integrity of voice transmissions.

Voice-enabling the data network

Unlike most data applications, voice is very sensitive to delay. Good voice quality provides a faithful recreation of the conversation, with the inflection, pauses, and intonation used by the speakers. Long and variable delays between packets result in unnatural speech and interfere with the conversation. Dropped packets result in clipped speech and poor voice quality. Fax transmissions are even more sensitive to the quality of the transmission and are less tolerant of dropped packets than voice.

One way to deal with the problem of delay and congestion is to add bandwidth to the network at critical junctures. Although this is feasible in the backbone, it is a costly and ineffective solution in the access arena, defeating the "bandwidth sharing" benefits of packet networks. The best solution is to implement mechanisms at the customer premises, access node, and backbone that manage congestion and delay—without increasing bandwidth—such as setting priorities for different types of traffic. Therefore, smart access equipment was

developed that could implement procedures to reduce network congestion and the delay of voice packets without adding bandwidth.

VoFR technology

The significance of VoFR is very important because of its wide existence in the data network infrastructure. To have a flawless converged network with Frame Relay network, we have to ensure several requirements of how VoFR can be deployed. Of the three popular packet/cell technologies (Frame Relay, IP, and ATM), Frame Relay is the most widely deployed. It is commonly used in corporate data networks due to its flexible bandwidth, widespread accessibility, and support of a diverse traffic mix and technological maturity.

Frame Relay service is based on PVCs. This service is appropriate for closed user groups and is also recommended for star topologies and predictable performance needs. VoFR is a logical progression for corporations already running data over Frame Relay.

VFRADs integrate voice into the data network by connecting the router, SNA controller, and the PBX at each site in the corporate network to the Frame Relay network.

Many VFRADs employ sophisticated techniques to overcome the limitations of transporting voice over the Frame Relay network without the need to add costly bandwidth.

VoFR prioritization

As we know, Voice is a real-time application. Prioritization over any other packet is mandatory to have a successful data/voice converged network. The VFRADs' prioritization schemes "tag" different applications according to their sensitivity to delay, assigning higher priority to voice and other time-sensitive data such as SNA. The VFRADs let the higher-priority voice packets go first, keeping the data packets waiting. This has no negative effect on data traffic, as voice transmissions are relatively short and, being compressed, require very little bandwidth. They can therefore slip into the data network alongside the heavy graphics, payroll information, e-mail, and other data traffic without perceptibly encumbering overall network performance.

Frame Relay service providers have also begun to offer different QoS. Users can purchase the highest QoS, Real-Time Variable Frame Rate, for voice and SNA traffic. The second quality Frame Relay service, Non-Real Time Variable Frame, is typically purchased for LAN-to-LAN and business-class Internet and intranet traffic. The lowest QoS, Available/Unspecified Frame Rate, is used for e-mail, file transfer, and residential Internet traffic. In addition, the VFRAD can be configured to assign less sensitive traffic with a Discard Eligibility (DE) bit. These frames will be dropped first in case of network congestion.

Fragmentation

VFRADs incorporate fragmentation schemes to improve performance. Data packets are divided into small fragments, allowing higher-priority voice packets to receive the right-of-way without waiting for the end of long data transmissions. The remaining data packets in the data stream are momentarily halted until the voice transmission gets through (see Figure 3.2-5).

The downside of fragmentation is that it increases the number of data frames, thereby increasing the number of flags and headers. This also increases overhead as well as reduces bandwidth efficiency. RAD's FR+ application provides an enhanced fragmentation mechanism that fragments data frames only in cases where voice packets arrive at the switch during a data transmission. Otherwise, the long data frames are sent intact.

Controlling variable delay on VoFR

Variation in the arrival times between voice or data packets, also called jitter, causes unnatural-sounding voice instead of a smooth voice stream. If a packet does not arrive in time to fit into the voice stream, the

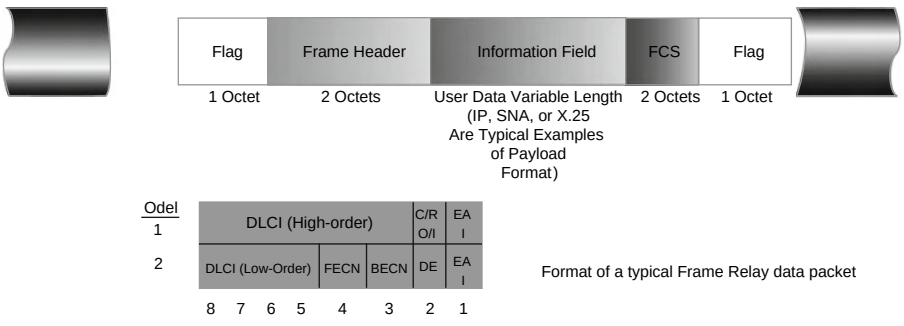


Figure 3.2-5 A typical Frame Relay data packet.

previous packet is replayed. This can seriously detract from voice quality. To avoid the effect of jitter, VFRADs detain each packet in a jitter buffer, giving subsequent packets time to arrive and still fit into a natural voice flow. Because the jitter buffer adds to the overall delay of voice transmissions, the optimal jitter buffer should fit the network's differential delay. Also some advance access devices in the market employ adaptive jitter buffering, which continuously monitors the network delay and adjusts the queuing period accordingly.

Voice compression

Voice compression allows the packet switching network to most effectively carry a combination of voice and data sessions without compromising voice quality. Because Frame Relay access is usually at data rates of 56/64 kbps, low bit-rate voice compression algorithms such as ITU G.723.1 and G.729A permit the greatest number of simultaneous multiple calls while maintaining high-quality voice. Many vendors have implemented voice compression algorithms in their Frame Relay access devices, and can offer greater bandwidth savings, reduced network congestion, and high-quality voice transmissions.

Silence suppression on VoFR

In a telephone conversation, only about 50 percent of the full duplex connection is used at any given time. This is because, generally, only one person talks while the other person listens. In addition, voice packets are not sent during interword pauses and natural pauses in the conversation, reducing the required bandwidth by another 10 percent. Silence suppression frees this 60 percent of bandwidth on the full duplex link for other voice or data transmissions.

Echo cancellation on VoFR

Echo cancellation improves the quality of voice transmissions. It eliminates the echo that results from the reflection of the telephony signal back to the caller, which can occur in a four-wire to two-wire hybrid connection between the VFRAD and the telephones or PBX. The longer it takes the signals to return to the caller, the more perceptible the echo.

Voice switching requirements

Various vendors implement voice switching in their VFRADs in different ways. Users should consider the requirements listed following when examining the voice switching capabilities of vendor equipment:

- Flexible numbering plan
- Transparent integration of voice and data
- Easy setup
- Minimal number of virtual links (DLCIs) between sites
- Voice, signaling, and data over a single DLCI
- Efficient switching of compressed voice
- Support and conversion of signaling methods
- Transparent physical interfaces
- Integrated fax support

In addition, the VFRAD should offer the following advanced PBX features:

- Programmable out-of-service indication
- Digit manipulation and DTMF generation (storing and forwarding)
- Hunt Groups

VoFR devices compliant with implementation agreement are not required to negotiate operational parameters. Negotiation procedures are for further study. Therefore, at the time of provisioning, the network manager must configure end-to-end configuration parameters (e.g., Vocoder). End-point devices providing the VoFR service are configured with compatible sub-channel assignments, signaling, compression algorithms, and other options. The relationship of the VoFR service, VoFR Service user, and the frame relay service is shown in Figure 3.2-6.

FRF. 11 enables vendors to develop standards-based equipment and services that interoperate. It also enables network managers seeking to reduce communications costs and maximize their frame relay network to consider VoFR as an option to standard voice services. In some cases, users may find they have excess bandwidth in their frame relay network that could efficiently support voice traffic. Other telecommunications managers may find that the incremental cost of additional frame relay bandwidth for voice traffic may be more cost-effective than standard voice services offered by local or long-distance carriers.

VoFR can provide end users with a cost-effective option for voice traffic transport needs between company locations. For instance, the network manager may integrate some voice channels and serial data over a frame relay connection between a branch office and corporate headquarters (see Figure 3.2-7). By combining the voice and data traffic on a frame relay connection already in place, the user has the potential to obtain cost-effective intracompany calling and efficient use of the network bandwidth.

There are potential trade-offs when implementing VoFR. These include

- loss of the quality commonly associated with toll traffic due to VoFR's use of voice compression
- loss of management and administrative benefits associated with carrier voice services (i.e., the loss of consolidated voice billing and invoice itemization, end

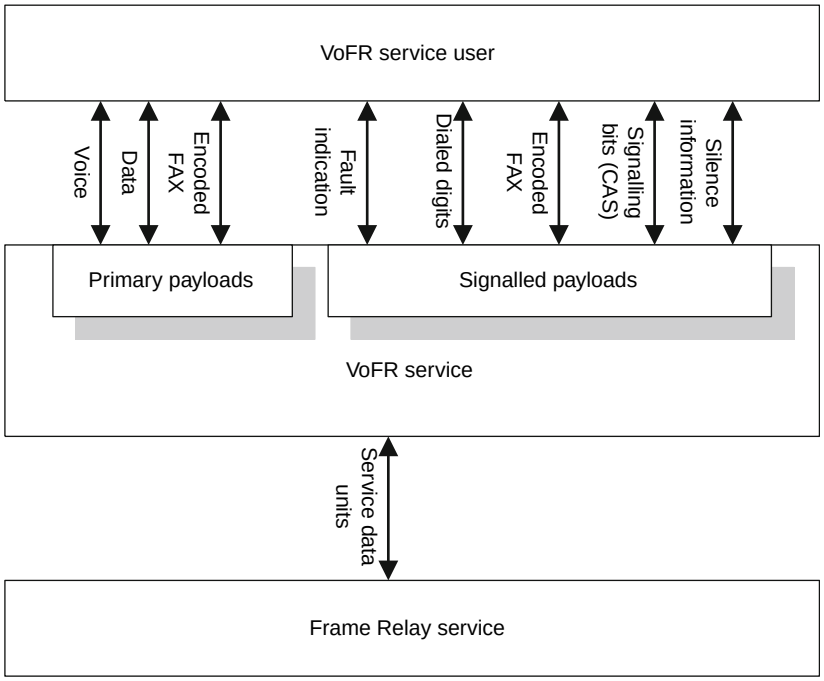


Figure 3.2-6 Layered architecture of VoFR.

user charge back capabilities, and other advanced features such as ID and accounting codes)
- lack of equipment interoperability between the customer premise and the equipment vendor
- lack of standards defining the acceptable levels of quality for voice transport over a carriers' frame relay network.

These trade-offs do not necessarily negate the value and promise of VoFR. Significant advances in digital signal processors and compression algorithms often provide voice at a level approaching toll quality, for a fraction of the cost of public service. VoFR vendors continue to add advanced capabilities in management and administration capabilities. In addition, future industry work will also seek to develop standards which define acceptable levels of quality and performance metrics for voice transport through carriers' frame relay networks.

Voice over Internet Protocol

One of the greatest challenges for VoIP is to develop networks that are not only scaleable but also seamless to the subscriber and to the service provider. If the service is difficult to access by the subscriber due to complex dialing plans and special PIN numbers, or requires significant time to complete a call, or has constant call drops, then the IP gateway will only be used by a limited client base. VoIP is not a new network, but a new application on IP networks. Traditionally, the voice is transported on a network that uses circuit switching technology, where data networks are built packet-switched technology. There are various reasons this transition is taking place, many of which have to do with economies of scale. Traditionally a Telephony network has been architected around circuit switch technology, requiring specific equipment and management techniques. Networks have emerged from being difficult to implement, a side thought for many companies to a critical part of their business strategy and an integral part of their economic growth.

The overall advantage of VoIP comes from treating voice as another form of data. While claims that the PSTN is dying are premature and unfounded, the advantages presented by IP telephony are clearly visible today.

More bandwidth

One advantage of VoIP is that it dramatically improves the efficiency of bandwidth for real-time voice transmission, in many cases by a factor of 6 or more. This increase in efficiency is a real long-term driver for the evolution from circuit-switched to packet-switched technology.

New services

Another advantage IP telephony has over the PSTN is that it enables the creation of a new class of service that

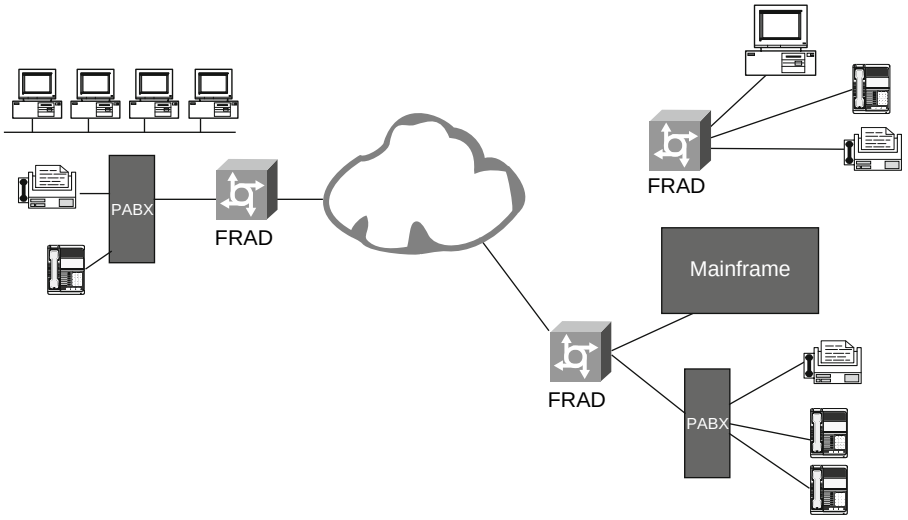


Figure 3.2-7 Integrated Voice and Data over Frame Relay network.

combines the best characteristics of real-time voice communications and data processing, such as web-enabled call centers, collaborative whiteboarding, multimedia, telecommuting, and distance learning. This combination of human interaction and the power and efficiency of computers is opening up an entirely new world of communications.

Progressive deployment

The final advantage of VoIP is that it is *additive* to today's communications networks. IP telephony can be used in conjunction with existing PSTN switches, leased and dial-up lines, PBXs and other customer premise equipment (CPE), enterprise LANs, and Internet connections. IP telephony applications can be implemented through dedicated gateways that in turn can be based on open standards platforms for reliability and scalability.

Market size

There is a wide range of numbers describing the current size of the IP telephony market and the growth of the market over the next three to five years. While the specific projections vary, even the most conservative analysts are predicting phenomenal growth. The numbers are summarized in Figure 3.2-8.

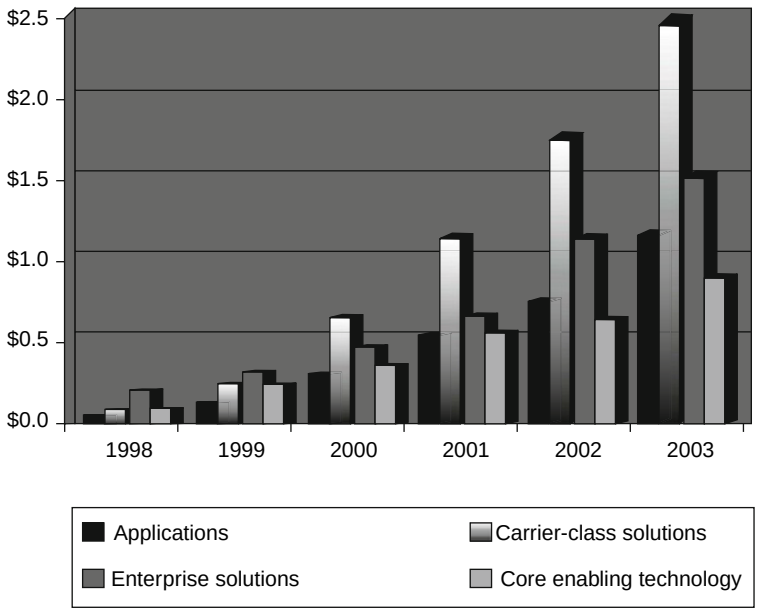


Figure 3.2-8 IP telephony growth.

VoIP is very complex because it involves components of both the data and the voice world. VoIP is a revolutionary solution as, historically, data and voice worlds have used two different networks, two different support organizations, and two different philosophies. The voice network has always been separate from the data network because the characteristics of voice applications are very different from the characteristics of data applications. The traditional voice network is circuit switched. Interactive voice traffic is sensitive to delay and jitter but can tolerate some packet loss. The voice philosophy was to ensure the "five nines" of reliability—99.999 percent—because the lack of communication might be life threatening (i.e., the inability of placing a "911" call for help). Voice calls have always been networked through dedicated lines to maintain no delay and reliability. The data network, on the other hand, is packet switched. Data is less sensitive to delay and jitter, but cannot tolerate loss. The data philosophy has been concerned with providing reliable data transmission over unreliable media, regardless of delay. Bandwidth in the data world is largely shared, so congestion and delay are often present for multimedia applications such as voice.

Voice and Data network processes work very differently. Data can tolerate delay or latency at a certain limitation, but voice application is very sensitive to delay, latency jittering, and packet loss, and requires much more attention than a data packet. Figure 3.2-9 shows a typical VoIP converged network with different type media and signaling gateway.

Implementing VoIP requires attention to many factors, including

Delay
Packet loss
Available bandwidth
Manageability
Jitter
Packet prioritization
Latency
Qos

Table 3.2-5	
Media gateway controller	Coordinates setup, handling, and termination of media flows at the media gateway.
Signaling gateway	SS7-IP interface, coordinates SS7 view of IP elements and IP view of SS7 elements.
Media gateway	Terminates PSTN lines and packetizes media streams for IP transport.

Delay

In VoIP excessive end-to-end delay makes conversation inconvenient and unnatural. Each component in the transmission path—sender, network, and receiver—adds delay. ITU-TG.114 (One-Way Transmission Time) recommends 150 ms as the maximum desired one-way latency to achieve high-quality voice. Packet delay is the length of time it takes a packet to traverse the network. Users will experience difficulties in carrying on a normal conversation when the one-way network delay exceeds 50 ms. Packet delay in excess of 50 ms can have a noticeable effect. However, some applications or users may elect to tolerate it, just as many people accept substandard quality when using cell phones.

Packet loss

Typically occurs either in bursts or periodically due to a consistently congested network. Periodic loss in excess of 5–10 percent of all voice packets transmitted can degrade voice quality significantly. Occasional bursts of packet loss can also make conversation difficult. In a VoIP network, there are several factors influencing packet loss.

- Packet loss requirements are tighter for tones (other than DTMF) than for voice. The ear is less able to detect packet loss during speech (variable pitch) than it is during a tone (consistent pitch).

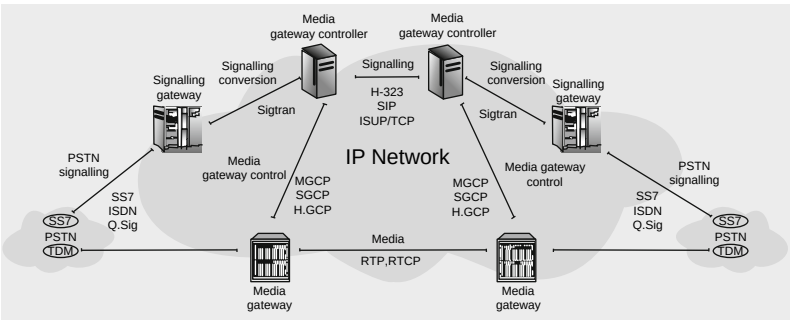


Figure 3.2-9 Converged architecture.

Table 3.2-6 Sample delay budget table.

Parameter	Fixed delay	Variable delay
CODEC (G.729)	25 msec	
Packetization	Included in CODEC	
Queuing delay		Depends on uplink. In the order of a few mSec.
Network delay	50 msec	Depends on network load.
Jitter buffer	50 msec	
Total	125 msec	

- Packet loss requirements are tighter for short, continuous packet loss than for random packet loss over time. Losing ten contiguous packets is worse than losing ten packets evenly spaced over an hour timespan.
- Packet loss may be more noticeable for larger voice payloads than for smaller ones because more voice is lost in a larger payload.

Available bandwidth

IP network transmission adds considerable overhead to any given data stream. Each protocol layer adds header and/or trailer information to be transmitted. Depending on their design, IP networks may show hard bandwidth limits or soft limits in which delays and/or losses increase as traffic load increases and congestion builds. For audio traffic, three main approaches are available to minimize bandwidth usage, as follows:

1. choose a low-bit-rate audio codec;

2. combine multiple audio frames into one packet;

3. suppress transmission of silence.

The first method can reduce the audio data rate from 64 kbps (PCM) to 8 kbps (G.729), or even as low as 5.6 kbps (G.723.1). (Note that addition of RTP/IP overhead makes the net traffic differences smaller than the data rate differences, though the use of header compression can mitigate this effect.) Two sacrifices must be made to obtain this reduction: Processing must be expended to perform the encoding and decoding; and, the audio quality is degraded to an extent dependent on the codec used. The second method reduces packet overhead, does not affect audio reproduction, but does increase the latency by the time represented by the number of additional frames in each packet. The third method can reduce overall audio traffic by roughly half, but the success of this effort depends on the statistical averaging effects of many calls. Another approach to bandwidth management is through prioritization: Give audio preferred access to the limited facilities and let other types of data streams contend for what's left. This approach may be invoked by choice of an 802. 1p priority (at layer 2), or a DiffServ service tag (at layer 3).

Despite the high speeds available in today's LANs, they are not immune to QoS issues. Shared media, for example, can cause problems if collisions occur (because at least one user may need to be delayed). Prioritization of traffic on the LAN allows QoS requirements to be signaled to LAN switches and routers. The IEEE ratified a standard for QoS prioritization in IEEE802 LANs in 1999. The IEEE802.1p specification defines three bits within the IEEE802.1Q field in the MAC header (which is a part of OSI Layer 2). The IEEE802.1Q field was initially designed to support VLAN interoperability but has been extended to support traffic priorities. This information can be used as a "signal" to any device that can decode the bits. Figure 3.2-10 illustrates the approach.

Three bits allows up to eight settings that can be used for classes of traffic and priorities. Typically, an NIC card in a LAN-attached system sets the bits according to its needs, and this information can be used by Layer 2 switches to direct the forwarding processes. If multiple distinct LANs are interconnected via routers (i.e., Layer 3 switches), then the Layer 2 bits must be used to drive Layer 3 QoS mechanisms. Although the IEEE802.1p/Q mechanism cannot operate on an end-to-end basis in an internetwork, it does provide a relatively simple method of defining and signaling an end system's requirements within a LAN environment. Multimedia applications and convergence changed the situation—usage and users must now be distinguishable and network processing needs to be tailored to each class of traffic. For example, a real-time telephone call should not be handled the same way as a large file transfer if they share a link. Differentiating among QoS requirements can be done by classifying each packet and using this information in switching decisions. The IETF (in a working group called DiffServ) is completing a series of standards that re-define the type of service (ToS) byte and re-name it as the Differentiated Services Control Byte (DSCP). RFCs 2474 and 2475 provide the details. Figure 3.2-11 illustrates the DSCP within the context of the IP protocol header.

DiffServ [RFC2474] makes use of the existing Type-of-Service (ToS) octet in the existing IP Version 4 header [RFC791]. As such, it may be set by information senders and used by IP (Layer 3) routers within the enterprise network. Effectively, the ToS octet is really just 6 bits wide because neither DiffServ nor IP ToS use the least-significant 2 bits; they are reserved for future use and fixed at 00. The original ToS definition was: bits 0–2 Precedence (7 = highest, through 0 = lowest), bit 3

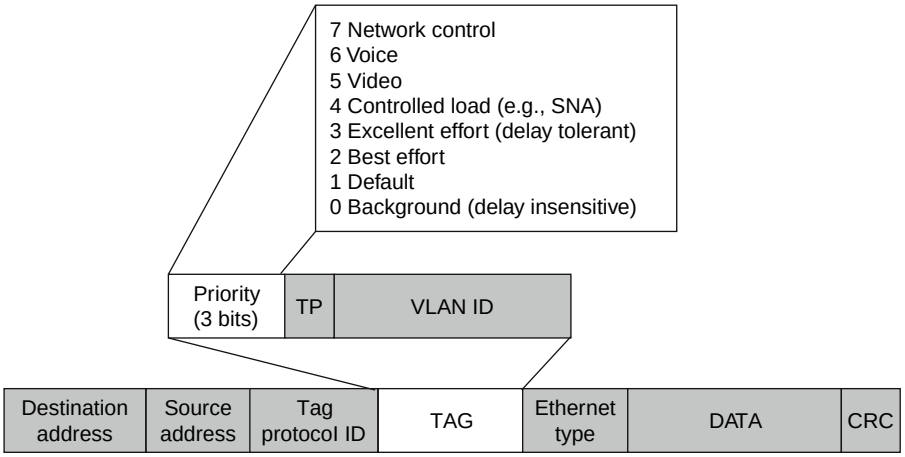


Figure 3.2-10 IEEE802.1p/Q prioritization bits.

Delay (0 = normal delay, 1 = low delay), bit 4 Throughput (0 = low, 1 = high), bit 5 Reliability (0 = normal, 1 = high), bits 6–7 reserved for future use (set to 00) DiffServ combines bits 0–5 into a single selector for "Per-Hop Behavior" (PHB).

Compatibility with older systems using precedence is preserved by fixing the 8 PHB values of the form xxx000 (binary) to be equivalent to the behaviors provided by networks that provided precedence routing; in particular, the value 000000 must always represent the default "best effort" service provided by each IP network. Generally, the 6 PHB bits represent 64 "code points" or possible per-hop behaviors. These 64 are divided into three sets: xxxxx0 for (future) standardization; xxxx01 and xxxx11 for experimental or local use. Not all values may be supported by all the networks, and it may be necessary to map the PHB values from one network to another to obtain a relatively consistent behavior. Until PHB values are standardized across networks, and all networks implement DiffServ, each endpoint on a network must be configured to use the value(s) appropriate for that network's traffic, and let the inter-network gateways translate, if required. In the absence of any prior knowledge, a participant in an IP conversation could start with the default PHB codepoint (000000), then move to another when it discovered a different codepoint being received with the same type of traffic from the other party to the conversation. Once the codepoint is changed from 000000, however, this procedure should be abandoned to avoid constant re-mapping if the intervening networks do not have a strictly reversible mapping.

Manageability

On a converged network, to have a seamless performance it is crucial to manage your network carefully and

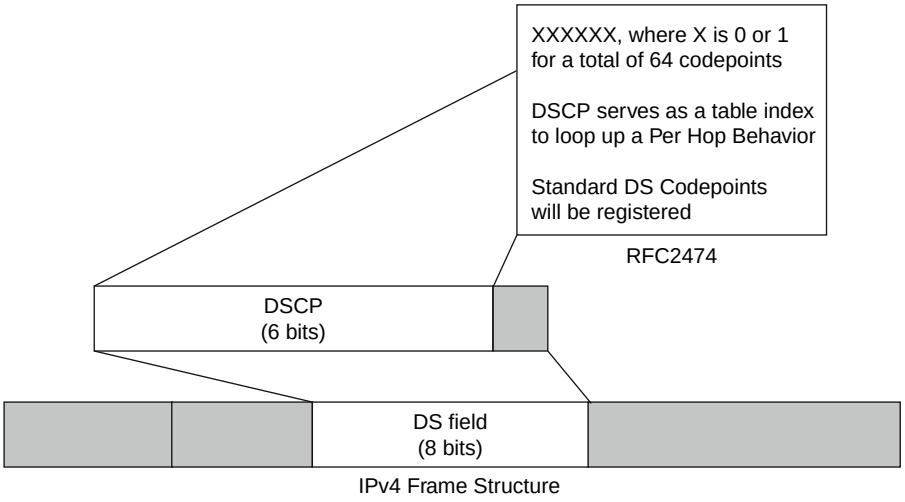


Figure 3.2-11 Differentiated services bits.

efficiently manage your limited bandwidth, particularly for real-time applications such as voice packets over the network. One of the options to manage your network in an efficient manner is a policy-based management tool with a real-time voice/data monitoring tool. Usually an excellent class policy based with real-time voice/data and converged network management system supports configuring the network in accordance with business needs and requirements. Most network vendors offer or support PC/Windows applications that provide a means to express business policies in terms of rules that govern the behavior of the network. These rules may, in turn, be expressed in terms of users, applications, the time of day, and monitor real-time applications. Policy-based network rules configure (and reconfigure from time to time, as necessary) the network so that its behavior carries out the business needs. It implements the rules by configuring individual network devices to behave in accordance with the network policies, typically utilizes a variety of protocols, such as telnet/Command Line Interface, COPS, and SNMP, to communicate with network devices to achieve the desired configuration. The way policy-based network management works, the rules are generally stored in a directory with a schema designed to support rule-based policy management. Each IP switch and router in the converged enterprise network should be able to be configured to provide a variety of services. These services provide low latency and constant delay as is required for good voice quality. This service is generally configured by assigning a particular queue(s) in the device that is to be used for this specific traffic. This traffic may be designated as that originating from one or more specific physical ports on the device, that originated from a UDP port within a specified range, or as that tagged with specific 802.1p/Q and/or DiffServ code point values. The assigned queue(s) is(are) then given priority over other queues that are used to forward latency-insensitive traffic. Though all of these mechanisms may be utilized under control of the policy-based management system, only UDP port ranges and packet tagging via 802.1p/Q and DiffServ are flexible enough to be used entirely under software control interests and are the services that provide low latency and constant delay as is required for good voice quality.

Jitter

Perhaps the one aspect of IP networks that makes them fundamentally different from circuit-switched networks is variance in propagation delay, or jitter. Audio encoding/decoding is basically a synchronous process: The analog signal is sampled at prescribed intervals, the samples (or their encoding) are transmitted, and the received samples are clocked out to reproduce the analog signal. IP networks are basically asynchronous transport devices. Clocks at the sender and the receiver are not synchronized; hence, the receiver may play out audio slightly faster than it was generated (thereby sometimes running out of audio to play) or slightly slower (thereby falling further and further behind the source).

To make matters worse, individual packets of audio may experience more or less delay as they travel through the network. This causes variations in the inter-arrival time. Some packets may experience "infinite delay" (be lost) or such long delay that they become useless or stale. The main technique for dealing with variable propagation delay is the jitter buffer: Incoming audio is passed through the jitter buffer and some amount of audio information (PCM samples, say) is buffered up before playing the samples out. If an incoming packet is delayed somewhat, then the information already in the buffer can be played out until either the buffer empties or the information arrives. Because the packets are presumably coming at a fixed rate, when one or more packets arrive late, some subsequent packets should arrive "early" (or on time) to refill the buffer. If the buffer runs dry, some type of audio fill-in must be supplied. As network jitter gets worse, the size of the jitter buffer must be increased to avoid too many under-runs. Unfortunately, the use of a jitter buffer introduces a delay proportional to the buffer's target size. Some products (e.g., Microsoft NetMeeting) create an extremely large jitter buffer on the order of 200–300 ms in order to handle jitter from virtually any type of connection, which introduces a "fixed" delay degradation of the conversation.

Voice packet prioritization

Prioritization of network traffic is simple in concept: Give important network traffic precedence over unimportant network traffic. One prioritization scheme assigns priority based on the UDP (User Datagram Protocol) port numbers that the voice packets use. This scheme allows you to use network equipment that can mark the packets on these ports with a priority. UDP is used to transport voice through the LAN because, unlike TCP, it is not connection-based. Because of the human ear's sensitivity to delay, it is better to drop packets rather than retransmit voice in a real-time environment. Prioritization is also called CoS (class of service) because traffic is classed into categories such as high, medium, and low (gold, silver, and bronze), and the lower the priority, the more "drop eligible" is a packet. E-mail and Web traffic is often placed in the lowest categories. When the network gets busy, packets from the lowest categories are dropped first.

Prioritization/CoS should not be confused with QoS. It is a subset of QoS. A package-delivery service provides an analogy. You can request priority delivery for a package. The delivery service has different levels of priority

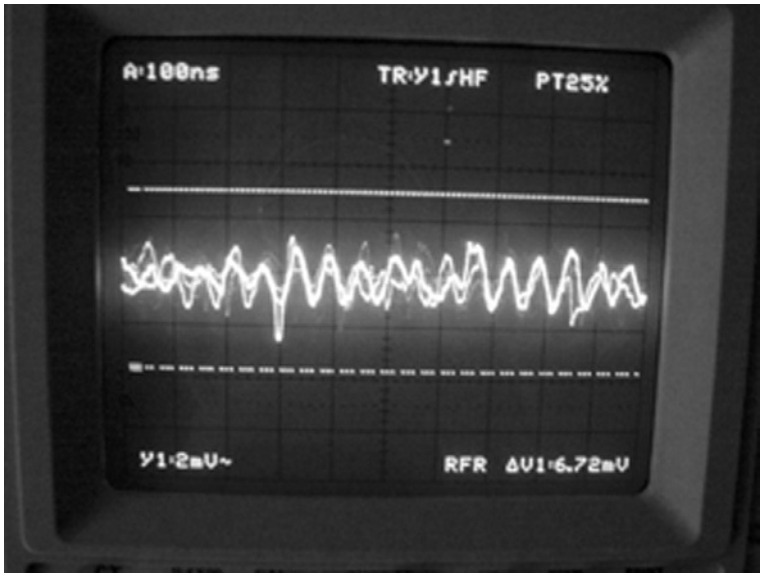


Figure 3.2-12 Sample jitter testing: What we see on the scope is not the signal that carries the jitter, but the jitter inside the signal.

(next day, two-day, and so on). However, prioritization does not guarantee the package will get there on time. It may only mean that the delivery service handles that package before handling others. To provide guaranteed delivery, various procedures, schedules, and delivery mechanisms must be in place. For example, Federal Express has its own fleet of planes and trucks, as well as a computerized package tracking system.

Prioritization has been used in multi-protocol routers to give some protocols higher priority than other protocols. For example, SNA (Systems Network Architecture) traffic will time-out if it is not delivered promptly, causing retransmissions that degrade network performance. Such protocols should be given high priority. A number of other prioritization/CoS schemes are outlined here.

- MAC layer prioritization. In a shared LAN environment such as Ethernet, multiple stations may contend for access to the network. Access is based on first-come, first-serve. Two stations may attempt simultaneous access, causing both stations to back off and wait before making another attempt. This is minimized for switched Ethernet where only one station is connected to a switch port. A number of vendor-specific Ethernet priority schemes have been developed. Token ring networks have a priority mechanism in which a reservation bit is set in tokens to indicate priority.
- VLAN tagging and 802. 1p The IEEE 802.1Q frame-tagging scheme defines a method for inserting a tag into an IEEE MAC-layer frame that defines membership in a virtual LAN. Three bits within the tag define eight different priority levels. The bit settings serve as a label that provides a signal to network devices as to the class of service that the frame should receive.
- Network layer prioritization The IP packet header has a field called ToS. This field has recently been redefined to work with the IETF's Differentiated Services (Diff-Serv) strategy. Diff-Serv classifies and marks packets so that they receive a specific per-hop forwarding at network devices along a route. The ToS bit is set once, based on policy information, and then read by network devices. Because IP is an internet-working protocol, Diff-Serv works across networks, including carrier and service provider networks that support the service. Therefore, Diff-Serv will support CoS on the Internet, extranets, and intranets.

Priority settings may be made in several places. The most logical place is the application running in the end user's system. But applications may not support the various schemes that are available, meaning that edge switches may need to infer priority levels for frames or packets by examining the contents of the packets. This is now easily done with so-called "multilayer switches" based on policies that are defined in policy-based management systems.

Latency

The average time delay experienced between an audio source (speaker) and destination (listener) does not affect the quality of speech directly, but it affects the speaker's perceptions of echo from the listener's end, and it affects the quality of a conversation (the ability of the listener to "turn the speech path around" and

respond). The most practical means of treating the echo problem is to cancel it at the endpoint closest to the cause of the echo. For an IP endpoint (e.g., telephone set), this means within the set itself. Endpoints with speakers and microphones can be troubled by acoustic echo. When an IP connection is inter-worked to the circuit-switched network, it is necessary to cancel any hybrid-induced echo that might be generated from a circuit-switched endpoint. This cancellation should take place in the interworking device. Telephone users expect good conversational quality on a call. Only in the most extreme cases (Earth to Moon) do they accept long delays without question. Frequent users of calls relayed via satellite have some acceptance of long delays, and are practiced in the required protocols, but most users are unwilling to accept such inconveniences on calls that are essentially "local." Although there is no specific breakdown point, the generally accepted standard is that one-way delays in excess of 150 ms or round-trip delays in excess of 300 ms are intolerable. Perhaps the most obvious contributor to latency is the audio packet propagation delay across the network (although it is often not the major contributor). Propagation delays (including media access delays) can vary from a few milliseconds on a LAN to hundreds of milliseconds and are strongly dependent on network loading and topology. The delays due to loading can be reduced by giving audio packets priority service over other network traffic. Various methods of prioritization are in use by vendors: explicit priority labeling (e.g., 802. 1p or DiffServ), or via address recognition (source/destination address preferences or port number preferences). Audio compression and packetization also contribute to latency. Most coders operate on an audio sample interval or frame, typically 10 ms. One or more encoded frames is then assembled into a packet for transmission (multi-frame packets reduce overhead). This represents a delay of one or more frame times. So it is crucial to monitor and control latency on your converged network.

Quality of service

Virtually all modern networks use a layered protocol structure with each higher layer taking on more responsibility for successful end system interworking. QoS mechanisms can be incorporated at any layer (or, in fact, in all of them). Table 3.2-7 illustrates the five layers of the TCP/IP stack and the positioning of the key QoS mechanisms.

Note that some of these mechanisms are inherent in the protocols (e.g., priority based on physical ports, decisions based on Layer 4 port numbers) rather than being explicitly added on for QoS control.

Most companies have become dependent on their networks. Who, for example, would be able to prosper in today's society without a telephone (even for calling someone in the same office)? Imagine the inconvenience if the telephone system was frequently so busy that you could not even get a dial tone or if a telephone call suddenly failed in the middle of a conversation? The same concerns arise when data networks are business critical, such is true for a bank teller or online book ordering. There can be little doubt that the quality of the network is very important, almost as important as quantity in some cases. The lack of adequate QoS can rapidly become a business inhibitor. Why then is there a sudden interest in dynamic QoS management for enterprise networks? Nothing about the traffic or its value has changed, although new forms of digitized traffic (such as video) have become available. What is different is that the underlying enterprise network technology is changing (to the TCP/IP architecture pioneered by the Internet) and the desire is now to "converge" all forms of traffic onto a single TCP/IP-based infrastructure. Convergence is aimed at reducing costs and complexity, increasing operational and managerial efficiency, and opening the door for innovative new multimedia applications. The original TCP and IP protocols were designed for applications such as e-mail distribution, basic file transfer, and remote terminal access, none of which were particularly time or bandwidth sensitive. The primary goal was to establish connectivity across a wide variety of systems and devices using a robust, inexpensive network. The philosophy was to assume that the network would do its best to deliver all messages that were submitted but with no guarantees of delivery or sequencing. This "best efforts" approach is quite satisfactory for basic data communications (and hence the popularity of the Internet) but has proved less suitable for networks that must also accommodate voice and video traffic (i.e., multiservice

Table 3.2-7

Application layer	User & application access, profile authorization, encryption, QoS-aw intelligent applications design for QoS
Transport layer	Port-based access control, TCP rate control
Network layer	Resource reservation, Type of Service bits, path controls, load balancing, address-based access control, protocol-based separation
Data link layer	Frame prioritization (802.1p/Q), ATM Cell QoS, VLAN and VPN isolation, logical port access control, path controls
Physical layer	Bit error correction, link diversity, path controls, physical security, physical port access

networking). The introduction of IP-based QoS mechanisms and the extension of network management systems is an essential step in the transition to the new, converged network infrastructure. Multiservice traffic is difficult to handle efficiently by very nature, especially in resource-limited environments.

The goal of the QoS strategy is to provide tools and mechanisms that allow the network administrator the means to optimize the performance of the network. This often involves tradeoffs between competing factors. For instance, basic PCM audio encoding is, by definition, a synchronous process; IP network transport is, by design, an asynchronous process. Traditional circuit switching provides dedicated, clocked resources for lockstep transmission; packet transmission is subject to unpredictable delays and losses. The network administrator's job is to use the QoS features to minimize the effects of IP network "unpredictability."

In a converged network, the voice call/feature server plays a central role in ensuring that the available QoS mechanisms are appropriately utilized so that voice traffic receives appropriate treatment in the network consistent with business policies regarding network utilization. This section describes the QoS mechanisms available in a converged network and how the call server controls their use. IP-based QoS mechanisms can be utilized to provide the best possible end-to-end audio experience when all or part of the audio path is carried over packet-switched networks. "Best," in this context, is defined by the customer as represented by the network administrator, and represents a tradeoff between audio reproduction quality, audio path delay (latency), audio loss, and network resource consumption. The network administrator is able to effect this tradeoff by means of audio codec selection and invocation of network prioritization through the Diff-Serv capability [RFC1889, RFC1890], and/or the IEEE 802.1p/Q MAC-layer prioritization and segregation. Use of Diff-Serv and/or 802.1p/Q priorities applies to either or both ends of a voice path: The voice call server configures media processor resources and the terminals involved in the call to use the appropriate values for the (sub)network(s) to which they are attached. Typically, these values will be uniform throughout the network, but allowances are made for different parts of the network to utilize different Diff-Serv and/or 802. 1p/Q values to designate the same service.

In order for voice traffic to receive excellent quality in a converged network, the network infrastructure equipment, i.e., industry standard IP switches and routers, and the voice endpoints, which include the telephones, softphones, and gateways, must be configured in a consistent manner so that voice packets are appropriately tagged and, consequently, properly identified and forwarded expeditiously. QoS architecture employs a policy-based network management system to consistently configure the network and the voice call/feature server with QoS parameter values. The voice call/feature server is, in turn, responsible for conveying these values to the voice endpoints.

This page intentionally left blank

Chapter 3.3

3.3

VoIP Technology

Juanita Ellis, Charles Pursell and Joy Rahman

In this chapter, we will discuss the unique issues that need to be considered when deploying voice across an IP-based network infrastructure. We will discuss the special needs that voice traffic has, such as minimal delay, jitter, and packet loss. Also, we will discuss the network protocols that can help support those needs, including 802.1p, Diffserv, and Priority Queuing. Finally we will provide some insight into some recommendations for the design of a converged network infrastructure.

Voice traffic needs in an IP network

Voice quality is always a subjective topic. Defining "good" voice quality varies with business needs, cultural differences, customer expectations, and hardware/software. The final determination of acceptable voice quality lies within the user's definition of quality and design, implementation and monitoring of the end-to-end data network.

Quality is not one discrete value where the low side is good and the high side is bad. There is a tradeoff between real-world limits and acceptable voice quality. Lower delay, jitter, and packet loss values can produce the best voice quality, but also may come with a cost to upgrade the network infrastructure to achieve the low network values. For example, in a Wide Area Network (WAN) link between Los Angeles and Manila, the link could add a large, fixed delay into the overall delay budget. This delay in the conversation may be noticeable to the end users as either an actual delay, or as echo, so the users may not get "toll quality" voice, but the cost savings can far override the slight loss of quality and the voice quality can still be acceptable for the user's purposes.

The concept of voice quality is also greatly influenced by the user's business needs. IP telephony quality can be engineered to several different levels to accommodate differing business needs. A small company may choose to implement IP telephony with acceptable sound instead of buying newer networking equipment to support excellent voice sound. A large call center company may want excellent voice sound as part of its corporate strategy, regardless of the cost of implementation.

The voice network has traditionally been separate from the data network because of the protocols used as well as the fact that the characteristics of voice applications are very different from those of data applications. Voice calls have had their own dedicated bandwidth throughout the circuit-switched network. This provided an environment where extremely high levels of reliability became the standard. Interactive voice traffic is sensitive to delay and jitter but can tolerate some packet loss, problems that were rarely an issue with circuit switching.

The data network, on the other hand, is packet switched. Data is less sensitive to delay and jitter, but cannot tolerate loss. The data philosophy has centered on providing reliable data transmission over unreliable media, almost regardless of delay. Bandwidth in the data world is largely shared, so congestion and delay are often present and can cause problems for multimedia applications such as voice.

The factors that affect the quality of data transmission are different from those affecting the quality of voice transmission. Data is generally not greatly affected by delay. Voice transmissions are degraded by relatively small amounts of delay and cannot be retransmitted. Additionally, a small amount of packet (data) loss does not affect voice quality at the receiver's ear, but even a small loss of data can corrupt an entire file or application. So in some cases, introducing VoIP to a high-performing data network can yield very poor voice quality.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2003 Elsevier Inc. All rights reserved

Implementing VoIP requires attention to many factors, including

- Delay
- Jitter
- Packet loss
- Packet mis-order
- Available bandwidth
- Packet prioritization
- Network design
- Endpoint audio characteristics (sound card, microphone, earpiece, etc.)
- Transcoding
- Echo
- Silence suppression
- Duplex
- Codec selection
- Router and data-switch configuration
- WAN protocols
- QoS (Quality of Service)/CoS (Class of Service) policy
- Encryption/Decryption

Delay

Packet delay is the length of time it takes a packet to traverse the network. Each element of the network—switches, routers, distance traveled through the network, firewalls and jitter buffers (such as those built into H.323 audio applications like Microsoft NetMeeting)—adds to packet delay. Router delay depends not only on hardware, but also on configurations such as access lists, queuing methods, and transmission modes. Delay (latency) can have a noticeable effect but can be controlled somewhat in a private environment (LAN/WAN) because the company or enterprise manages the network infrastructure or SLA (Service Level Agreement). When using the public network, there are inherent delays that one cannot control.

VoIP network vendors give many differing guidelines as to the amount of delay that is acceptable to maintain good voice quality. A good set of guidelines for voice quality is

- Under 150–200 ms delay can give very good voice quality.
- Delays exceeding 200 ms may still be quite acceptable depending on customer expectations, analog trunks used, codec type, etc.
- The H.323 protocol defines a maximum end-to-end delay of 400ms. Delays beyond this level can cause network instability.

  (These numbers are for delay between endpoints, meaning LAN/WAN measurements not including IP phones.)

The ITU-T has recommended 150 ms one-way delay (including endpoints) as the limit for "excellent" voice quality. One-way delays in excess of 250 ms can cause "talk-over," which is when each person starts to talk because the delay prevents them from realizing that the other person has already started talking.

Long WAN transports must be considered as a major contributor to the network delay budget. Some WAN service providers can lower delay in their network if it is negotiated and recorded as part of the company's SLA.

Jitter

Because network congestion can be encountered at any time within a network, buffers can fill instantaneously. This instantaneous buffer utilization can lead to a difference in delay times between packets in the same voice stream. This difference, called jitter, is the variation between when a packet is expected to arrive and when it actually is received.

Jitter can create audible voice-quality problems if the variation is greater than 20 ms (assuming an existing 20 ms packet size). Symptoms of excessive jitter are very similar to symptoms of high delay, because in both cases packets are discarded if the packet delay exceeds half the jitter buffer size.

To compensate for network jitter, many vendors implement a jitter buffer in their H.323 voice applications. The purpose of the jitter buffer is to hold incoming packets for a specified period of time before forwarding them to the decompression process. A jitter buffer is designed to smooth packet flow. In doing so, it can also add packet delay. Jitter buffers should be dynamic to give the best quality, or if static, should generally be sized to twice the largest statistical variance between packets. Router vendors have many queuing methods that alter the behavior of the jitter buffer. It is not enough to just select the right size of jitter buffer, one must also pair an appropriate queue-unloading algorithm type with the jitter buffer. The network topology can also affect jitter. Because there are fewer collisions on a data-switched network than on a hub-based network, there will be less jitter on the switched network.

Packet loss

Network packet loss occurs when packets are sent, but not received at the final destination due to some network problem. Qualifying problems caused by occasional packet loss are difficult to detect because each codec has its own packet loss concealment method. Therefore, it is possible that voice quality would be better using a compression codec (G.729A) compared to a full bandwidth

G.711 codec. Several factors make packet loss requirements somewhat variable, such as the following:

- Packet loss requirements are tighter for tones (other than DTMF) than for voice. The ear is less able to detect packet loss during speech (variable pitch), than during a tone (consistent pitch).
- Packet loss requirements are tighter for short, continuous packet loss than for random packet loss over time. Losing ten contiguous packets is worse than losing ten packets evenly spaced over an hour timespan.
- Packet loss may be more noticeable for larger voice payloads than for smaller ones, because more voice is lost in a larger payload.
- Packet loss may be more tolerable for one codec over another.
- Even small amounts of packet loss can greatly affect TTY (TDD) devices' ability to work properly.
- Packet loss for TCP signaling traffic increases substantially over 3 percent loss due to retransmissions.

Again, acceptable rates for packet loss vary with the needs of the end users.

- 1 percent or less can yield toll quality depending on many factors.
- 3 percent or less should give business communications quality.
- More than 3 percent may be acceptable for voice but may interfere with signaling.

Packet mis-order

Network packet mis-order is, for VoIP, very much like packet loss. If a packet arrives out of order, it is generally discarded as it makes no sense to play it out of order and buffers are small. Specifically, packets are discarded when they arrive later than the jitter buffer can hold them. Mis-order can occur when networks send individual packets over different routes. Planned events such as load balancing, unplanned events such as rerouting due to congestion, or other transient difficulties can cause packet mis-order. Packets traversing the network over different routes may arrive at their destination out of order. Network latency over multiple yet unequal routing paths can also force packet mis-order.

Transcoding

Transcoding is a voice signal converted from analog to digital or digital to analog (possibly with or without compression and decompression). If calls are routed using multiple voice coders, as in the case of call coverage on an intermediary system back to a centralized voice mail system, the calls may experience multiple transcoding (including the one in and out of the voice mailbox). Each transcoding episode results in some degradation of voice quality.

These problems may be minimized by the use of some form of network rerouting (e.g., Q-SIG Path Replacement). This feature detects that the call coming through the main call server has been routed from one tandem node, through the main, and back out to a third switch. In these cases, the system then reroutes the call directly, thus replacing the path through the main system with a more direct connection.

Echo

The two main types of echo are acoustic and impedance, although the sources of echo can be many. Echo will result when a VoIP call leaves the LAN through a poorly administered analog trunk into the PSTN. Another major cause is from an impedance mismatch between four-wire and two-wire systems. Echo also results when an impedance mismatch exists in the conversion between TDM (Time Division Multiplexing) networks and the LAN, or the impedance mismatch between a headset and its adapter. Impedance mismatch causes inefficient energy transfer. The energy imbalance must go somewhere and so it is reflected back in the form of an echo. Usually the speaker hears an echo but the receiver does not.

Echo cancellers, which have varying amounts of memory, compare the received voice with the current voice patterns. If the patterns match, the canceller cancels the echo. Echo cancellers aren't perfect, however. Under some circumstances, the echo gets past the canceller. The problem is exacerbated in VoIP systems. If the one-way trip delay between endpoints is larger than the echo canceller memory, the echo canceller won't ever find a pattern to cancel.

Silence suppression

Silence suppression software, often called Voice Activity Detection (VAD), monitors the received signal for voice activity. When no activity is detected for the configured period of time, the software informs the Packet Voice Protocol. This prevents the encoder output from being transported across the network when there is silence, resulting in additional bandwidth savings. The software also measures the idle noise characteristics of the telephony interface. It reports this information to the Packet Voice Protocol to relay this information to the remote end for comfort noise generation when no voice is present.

Some VADs can cause voice clipping and can result in poor voice quality, but the use of VAD can greatly conserve bandwidth and is therefore a very important detail

to consider when planning network bandwidth—especially in the WAN.

Duplex

The ideal LAN network for transporting VoIP traffic is a network that is fully switched from end to end because it significantly reduces or eliminates collisions. A network that has shared segments (hub-based) can result in lower voice quality due to excessive collisions and should be avoided.

Codec selection

Depending on the bandwidth availability and acceptable voice quality, it might be worthwhile to select a codec that produces compressed audio.

- A G.711 codec produces audio uncompressed to 64 kbps
- A G.729 codec produces audio compressed to 8 kbps
- A G.723 codec produces audio compressed to 5.3 to 6.3 kbps

Table 3.3-1 provides comparisons of several voice quality considerations associated with some of the codecs. These qualities are based on an industry guideline known as the Mean Opinion Score (MOS). The MOS is a subjective scale for rating voice quality from 1 to 5, where 4 or better is considered "toll quality" or the equivalent of the quality in the U.S. public switched telephone network (PSTN).

Generally, G.711 is used within LANs because bandwidth is abundant and inexpensive, whereas G.729 is used across WAN links because of the bandwidth savings and reliable voice quality.

PC Considerations using IP softphone

IP softphones are software on a PC that simulates a telephone. The "perceived" audio/voice quality at the PC endpoint is a function of at least four factors, as follows:

1. Transducer quality
   The selection of speaker and microphone or headset has an impact on the reproduction of the sound.
2. Sound card quality
   There are several parameters that affect sound card quality. The most important is whether or not the sound card supports full-duplex operation.
3. End-to-end delay
   A PC can be a major component of delay in a conversation. PC delay consists of the jitter buffer and sound system delays, as well as the number of other processes running and the speed of the processor.
4. Speech breakup
   Speech breakup may be the result of a number of factors.
   - Network jitter in excess of the jitter buffer size
   - Loss of packets (due to excessive delay, etc.)
   - Aggressiveness of silence suppression
   - Performance bottleneck in the PC

Lower-speed PCs (or PCs with slow hard drives) may have adverse interactions with sound playback and recording. This can cause breaks in received or transmitted audio. The best thing to do in this situation is to increase the processor speed, increase the amount of RAM, and/or reduce the number of applications competing for the processor or hard drive resources. One notable resource consumer is the Microsoft Find Fast program that launches from the Startup folder (and runs in the background). This application periodically re-indexes the hard drive and consumes significant PC resources in the process.

Table 3.3-1 Comparison of common speech coding standards.

Standard	Coding type	bit rate (Kbps)	MOS
G.711	PCM	64	4.3
G.729	CS-ACELP	8	4.0
G.723.1	ACELP MP-MLQ	6.3 5.3	3.8

Bandwidth requirements

The bandwidth available to the user is very important. Access to the network using slower connections, such as dial-up connections, will degrade voice quality. The best voice quality is achieved in both LANs and WANs when the bandwidth is "owned" by the customer. Customer-owned bandwidth can be shaped to optimize VoIP traffic. Conversely, bandwidth that is not controlled, like the Internet, cannot give consistent sound quality because it cannot be optimized for VoIP. Factors of delay, jitter, and packet loss are exacerbated over the public Internet, making the assurance of voice quality over the Internet especially problematic.

The bandwidth required for a given call varies with the voice codec used. Each codec converts the voice stream into a digital representation using varying amounts of bits to represent a second of voice. In addition to the codec's

Table 3.3-2 Typical VoIP packet.

Voice payload	RTP header	UDP header	IP header	Link header
X Bytes	12 Bytes	8 Bytes	20 Bytes	X Bytes

bit rate, the IP packet framing adds overhead to the data stream.

Properly provisioning the network bandwidth is a major component of designing a successful network. You can calculate the required bandwidth adding the bandwidth requirements for each major application (for example, voice, video, and data). This sum then represents the minimum bandwidth requirement for any given link. Many VoIP manufacturers recommend that minimum bandwidth should not exceed approximately 75 percent of the total available bandwidth for the link. This 75 percent rule assumes that some bandwidth is required for overhead traffic, such as routing and Layer 2 keepalives, as well as for additional applications such as e-mail and Hypertext Transfer Protocol (HTTP) traffic.

As illustrated in Table 3.3-2, a VoIP packet consists of the payload, IP header, User Datagram Protocol (UDP) header, Real-time Transport Protocol (RTP) header, and Layer 2 Link header. At a packetization rate of 20 ms, VoIP packets have a 160-byte payload for G.711 or a 20-byte payload for G.729. The IP header is 40 bytes, the UDP header is 8 bytes, and the RTP header is 12 bytes. The link header varies in size according to media.

The bandwidth consumed by VoIP streams is calculated by adding the packet payload and all headers (in bits), then multiplying by the packet rate per second.

This does not include Layer 2 header overhead and does not take into account any possible compression schemes, such as compressed Real-time Transport Protocol (cRTP).

Table 3.3-3 shows the typical bandwidth need for common codecs and sampling rates.

Table 3.3-3 Bandwidth consumption for voice payload only.

CODEC	Sampling rate	Voice payload in bytes	Bandwidth per conversation
G.711	20 ms	160	80 kbps
G.711	30 ms	240	53 kbps
G.729A	20 ms	20	24 kbps
G.729A	30 ms	30	16 kbps

Table 3.3-4 shows the total bandwidth, including layer 2 headers for various transmission media

These numbers are for each direction in a conversation, so a two-way conversation needs double this bandwidth across a given link.

There are methods to lessen some of these bandwidth needs. One is cRTP, which reduces the RTP header from 12 bytes per frame to 4. Another is to use silence suppression to suppress the packets sent by the non-speaking party in a conversation.

IP Network protocols to support voice

Voice quality is only as good as the quality of the weakest network link. Packet loss, delay, and delay variation all contribute to degraded voice quality. In addition, because network congestion (or more accurately, instantaneous buffer congestion) can occur at any time in any portion of the network, network quality is an end-to-end design issue. The QoS tools discussed here are a set of mechanisms to increase voice quality on data networks by decreasing dropped voice packets during times of network congestion and by minimizing both the fixed and variable delays encountered in a given voice connection.

Table 3.3-4 Bandwidth consumption with headers included.

CODEC	Ethernet 14 Bytes of header	PPP 6 bytes of header	ATM 53-byte cells with a 48-byte payload	Frame relay 4 bytes of header
G.711 (20 ms sample rate)	85.6 kbps	82.4 kbps	106 kbps	81.6 kbps
G.711 (30 ms)	156.5 kbps	54.4 kbps	70 kbps	54 kbps
G.729A(20 ms)	29.6 kbps	26.4 kbps	42.4 kbps	25.6 kbps
G.729A(30 ms)	19.5 kbps	17.4 kbps	28 kbps	17 kbps

All of these QoS schemes are discussed in more detail elsewhere. For our purposes here we will discuss how these QoS mechanisms can be used to optimize voice quality in a converged network.

In order for a VoIP solution to function well, the network must be able to give voice packets priority over ordinary data packets or sufficient bandwidth must always be available. There are several strategies available to provide this prioritization. These strategies include using CoS, prioritizing ports, prioritizing services, and using IEEE 802.1 p/Q to set the priority bits.

CoS versus QoS

The concepts of CoS and QoS are often confused. CoS is a classification method only. CoS does NOT ensure a level of QoS, but is the method used by queuing mechanisms to limit delay and other factors to improve QoS. Most CoS strategies assign a priority level, usually 0–7 or 0–63, to a frame or packet respectively. Common CoS models include the IP TOS (Type Of Service) byte, Differentiated Services Code Point (DiffServ or DSCP, defined in RFC 2474 and others), and the IEEE 802.1 p/Q.

QoS involves giving preferential treatment through queuing, bandwidth reservation, or other methods based on attributes of the packet, such as CoS priority. A service quality is then negotiated. Examples of QoS are CBWFQ (Class-Based Weighted Fair Queuing), RSVP (Resource Reservation Protocol—RFC 2205), MPLS (Multi Protocol Label Switching—RFC 1117 and others).

CoS, or tagging, is totally ineffective in the absence of QoS because it can only mark data. QoS relies on tags or filters to give priority to data streams.

UDP ports

One prioritization scheme assigns priority based on the UDP (User Datagram Protocol) port numbers used by the voice packets. This scheme allows the use of network equipment to prioritize all packets from a port range. UDP is used to transport voice through the LAN because, unlike TCP, it is not connection based. Because of the human ear's sensitivity to delay, it is better to drop packets rather than retransmit voice in a real-time environment, so a connectionless protocol is preferable to a connection-based protocol.

VoIP devices and servers can define a port range for voice priority. Routers and layer 3 data switches can then use these ports to distinguish priority traffic. This priority traffic can be voice packets (UDP), signaling packets (TCP), or both. This is an OSI model layer 4 solution and works on data coming to and from the specified ports or port range.

Differential services (DiffServ, DSCP, TOS)

The Differential Services Code Point (DSCP) prioritization scheme redefines the existing ToS byte in the IP header by combining the first 6 bits into 64 possible combinations. This use of the ToS byte is still evolving but can be used now by vop media controllers, IP telephones, and other network elements such as routers and switches in the LAN and WAN.

Please note that older routers may require a DSCP setting of 40 (101,000), which is backward compatible to the original ToS byte definition of critical. The ToS byte is an OSI model layer 3 solution and works on IP packets on the LAN and possibly the WAN depending on the service provider.

IEEE 802.1 p/Q

Yet another prioritization scheme is the IEEE 802.1Q standard, which uses four bytes to augment the layer-2 header. IEEE 802.1Q defines the open standard for VLAN tagging. Two bytes house twelve bits used to tag each frame with a VLAN identification number. The IEEE 802. 1p standard uses three of the remaining bits in the 802.1Q header to assign one of eight different classes of service. IEEE 802.1Q are OSI layer-2 solutions and work on frames.

VLANs

VLANs provide limited security and create smaller broadcast domains through software by creating virtually separated subnets. Broadcasts are a natural occurrence in data networks from protocols used by PCs, servers, switches, routers, NOS, etc. Creating a separate VLAN for voice reduces the amount of broadcast traffic (and unicast traffic on a shared LAN) the telephone will receive. Separate VLANs result in more effective bandwidth utilization and reduce the processor burden on the IP telephones and PCs by freeing them from having to analyze irrelevant broadcast packets. VLANs, a layer-2 feature, are created in data switches. A voice VLAN can also be manually applied to an IP telephone or given by a DHCP server. CoS tagging and QoS policies can be applied at OSI layer 2 by using VLANs.

Separate voice and data VLANs are options that make sense for many users. Note, however, that VLAN implementation and maintenance can be substantial, and again, is an option even as a best practice.

Resource reservation protocol

Once a given packet has been labeled by one or more of the schemes already mentioned, there are two common methods of acting on those labels to give the voice packets priority. One method is to act on a per-packet basis by giving special treatment to the packets as they are processed. The second is to establish a specialized route for those packets to traverse in advance.

In this section we will talk about the most common protocol used to support that second method, the RSVP.

RSVP is a signaled QoS configuration mechanism. It is a protocol by which applications can request end-to-end, per-conversation, QoS from the network and can indicate QoS requirements and capabilities to peer applications. RSVP is a layer-3 protocol suited primarily for use with IP traffic.

There are two significant RSVP messages, PATH and RESV. Transmitting applications send PATH messages toward receivers. These messages describe the data that will be transmitted and follow the path that the data will take. Receivers send RESV messages. These follow the path seeded by the PATH messages, back toward the senders, indicating the profile of traffic that particular receivers are interested in. In the case of multicast traffic flows, RESV messages from multiple receivers are "merged," making RSVP suitable for QoS with multicast traffic.

As defined today, RSVP messages carry the following information:

- How the network can identify traffic on a conversation (classification informaton)
- Quantitative parameters describing the traffic on the conversation (data rate, etc.)
- The service type required from the network for the conversation's traffic
- Policy information (identifying the user requesting resources for the traffic and the application to which it corresponds).

How RSVP works

PATH messages wind their way through all network devices en route from sender to receivers. RSVP-aware devices in the data path note the messages and establish state for the flow described by the message.

When a PATH message arrives at a receiver, the receiver responds with a RESV message (if the receiving application is interested in the traffic flow offered by the sender). The RESV message winds its way back toward the sender, following the path established by the incident PATH messages. As the RESV message progresses toward the sender, RSVP-aware devices verify that they have the resources necessary to meet the QoS requirements requested. If a device can accommodate the resource request, it installs classification state corresponding to the conversation and allocates resources for the conversation. The device then allows the RESV message to progress on toward the sender. If a device cannot accommodate the resource request, the RESV message is rejected and a rejection is sent back to the receiver.

In addition, RSVP-aware devices in the data path may extract policy information from PATH messages and/or RESV messages, for verification against network policies. Devices may reject resource requests based on the results of these policy checks by preventing the message from continuing on its path and sending a rejection message.

When requests are not rejected for either resource availability or policy reasons, the incident PATH message is carried from sender to receiver and a RESV message is carried in return. In this case, a *reservation* is said to be installed. An installed reservation indicates that RSVP-aware devices in the traffic path have committed the requested resources to the appropriate flow and are prepared to allocate these resources to traffic belonging to the flow. This process of approving or rejecting RSVP messages is known as *admission-control* and is a key QoS concept.

Guarantees must be valid *end-to-end* across multiple subnets. Lower quality guarantees can be provided without requiring tight coupling between the QoS mechanisms in different subnets. However, high quality guarantees require tight coupling between these mechanisms.

As an example, it is possible to independently configure devices in each subnet (in a top-down manner) to prioritize some set of traffic (as identified by IP port) above best-effort traffic (BBE service). This will indeed improve the quality of service perceived by the prioritized application, in all parts of the network. However, this is a low quality guarantee as it makes no specific commitments regarding available bandwidth or latency.

On the other hand, consider the quality of guarantee required to support a videoconference. A videoconferencing application requires that all subnets between the videoconferencing peers be able to provide a significant amount of bandwidth at a low latency. To do so efficiently requires that all devices along the data path commit the required amount of low latency bandwidth for the duration of the videoconference. As we have seen, high quality guarantees such as these generally require signaling across network devices in order to make efficient

use of network resources. In our sample network, multiple subnets, based on multiple media (and varying traffic handling mechanisms) must be coordinated via this signaling. RSVP is particularly suitable for this purpose because it expresses QoS requirements in high-level, abstract terms. Agents in each subnet are able to translate the media-independent, abstract requests into parameters that are meaningful to the specific subnet media.

Hosts generate RSVP signaling when it is necessary to obtain high quality guarantees. The network listens to this signaling at strategic points in the network. We will refer to devices that participate in RSVP signaling as *RSVP agents* or alternatively as *signaling* or *admission control agents*. As we have shown, appointing such agents at varying densities can provide varying quality/efficiency products. At a minimum we assume one or more admission control agents in each subnet. Each agent uses the mappings defined in ISSLL to translate high-level end-to-end RSVP requests into parameters that are meaningful to the media for which the agent is responsible. The admission control agent then determines, based on resource availability and/or policy decisions (with the cooperation of PDPs), whether an RSVP request is admissible or not. Any admission control agent along the route from sender to receiver may veto an RSVP request for resources. Requests that are not vetoed by any device are considered admitted and result in the return of an RSVP RESV message to the requesting transmitting host.

Prioritized queuing

As we discussed in the previous section, another common QoS mechanism is to give special treatment to packets marked as voice packets. This special treatment involves some form of prioritized queuing.

Prioritized queuing is a generic term for one of a number of schemes that allow one type of traffic to have priority over another when being transmitted over a bandwidth-limited media. The basic idea is for the queuing device (usually an IP switch or router) to analyze the packets, determine the type of payload the packet is carrying, compare that information to a set of prioritization and bandwidth utilization rules, and order the output according to those rules.

There are a number of different prioritization schemes used in currently available network devices today. The most commonly used schemes are known as Priority Queuing, Weighted Fair Queuing, and Class-Based Queuing. Many vendors use different names for these schemes and often combine the schemes together.

Priority queuing

In Priority Queuing, packets are forwarded on a strict priority basis. This means that packets in the higher-priority queue are sent first before any packets in the lower-priority queue (see Figure 3.3-1).

This works well for giving highly predictable performance for the high-priority packets, but lower-priority packets can be greatly delayed if there is significant high-priority traffic. Priority queuing works best in situations where the high-priority traffic has a relatively fixed bandwidth requirement (like voice). That allows priority queuing to be use while still giving predictable performance to the lower-priority traffic.

Weighted fair queuing

Weighted Fair Queuing (WFQ) is a method to overcome problems in strict Priority Queuing by giving all of the packets a weighting value based on priority and the anticipated time a given packet needs for processing. Once the packets are all given weights, they are transmitted in weight order. This gives the ability to still favor high-priority traffic, but give a fairer distribution of bandwidth to all other traffic types (see Figure 3.3-2).

The biggest problem with WFQ is that it is a very complex mechanism and therefore must be implemented in software, so the WFQ process can add delay to the overall transmission.

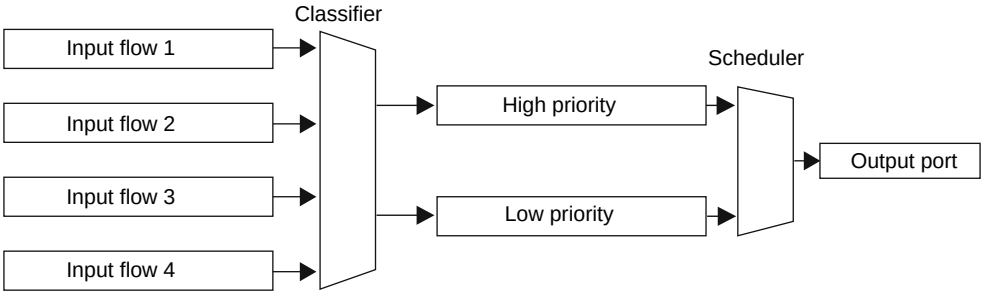


Figure 3.3-1 Priority queuing.

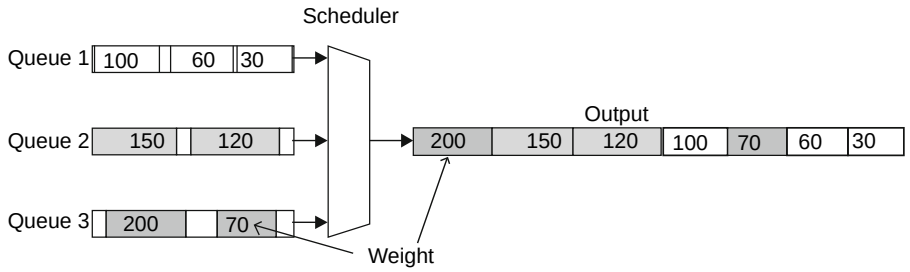


Figure 3.3-2 Weighted fair queuing.

Class-based queuing

Class-Based Queuing (CBQ) is a form of queuing where incoming packets are sorted into classes as in the other queuing schemes, each class is queued separately and assigned a fixed percentage of the output bandwidth. The output of the queues is then delivered in a given time period based on its percent allocation (see Figure 3.3-3).

This scheme allows all service types to have predictable performance. Also, due to the simpler nature of the queuing mechanisms, CBQ can be done in hardware, allowing for greater overall performance.

Most commercially available, converged network devices use one or more of these schemes to support simultaneous voice, data, and video traffic. The vast majority use either a CBQ scheme or a hybrid scheme where fixed bandwidth, high-priority traffic like voice or video conferencing are queued in a priority queue while the rest of the IP services use some form or WFQ or CBQ at the second level.

Other elements that affect VoIP

WAN considerations

Until WAN bandwidth becomes affordable at any speed, delivering bandwidth to applications over the WAN will remain a formidable task. When voice traffic is carried on packet networks, different labeling or queuing schemes function to give voice packets priority over data packets. The presence of large data packets may result in added serialization delay for VoIP packets across WAN links. This is due to the fact that smaller VoIP packets are held in queue while larger data packets are processed onto the WAN link. To avoid excessive delay, there may be benefit in fragmenting the larger data packets and interleaving them with the smaller voice packets. One technique is to adjust the packets by adjusting the Maximum Transmission Unit (MTU) size. Other techniques, such as Multilink point-to-point (MPP), Link Fragmenting and Interleaving (LFI) and Frame Relay Fragmentation (FRF 12), allow network managers to fragment larger packets as well as allow queuing mechanisms to speed the delivery of Real Time Protocol (RTP) traffic without significantly increasing protocol overhead or reducing data efficiency. Also, header compression protocols such as cRTP can and should be used between WAN links. Hardware-based cRTP is effective with very minimal delays, but software cRTP can add significant delay.

VPN (Virtual private network)

There are many definitions for Virtual Private Networks (VPN). For our purposes, VPNs refer to encrypted tunnels carrying packetized data between remote sites. VPNs can use private lines or use the Internet via one or more Internet Service Providers (ISP). VPNs are implemented in both dedicated hardware and software, but can also be integrated as an application to existing hardware and software packages. A common example of an integrated package is a firewall product that can

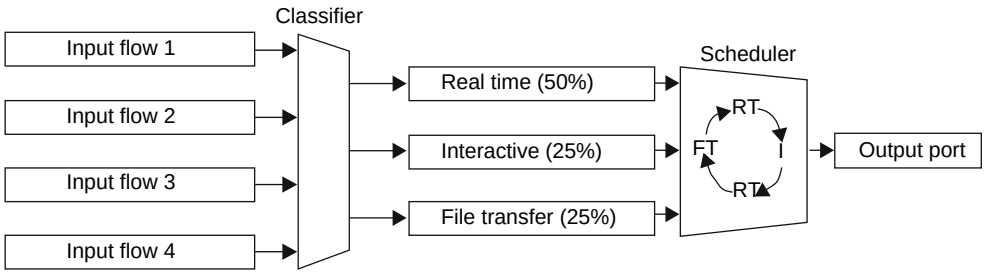


Figure 3.3-3 Class-based queuing.

provide a barrier against unauthorized intrusion as well as perform the security features needed for a VPN session.

The encryption process can take from less than 1 millisecond to 1 second or more, at each end. Obviously, VPNs can represent a significant source of delay and, therefore, negatively affect voice performance. Also, as most VPN traffic runs over the Internet and there is little control over QoS parameters for traffic crossing the Internet, voice quality may suffer due to excessive packet loss, delay, and jitter. Users may be able to negotiate a SLA with the VPN provider to guarantee an acceptable level of service.

Frame relay

The obstacle in running VoIP over frame relay involves the treatment of traffic within the Committed Information Rate (CIR) and outside of CIR, commonly termed the "burst range."

Traffic up to the CIR is guaranteed while traffic beyond the CIR typically is not. This is how frame relay is intended to work. CIR is a committed and reliable rate, whereas burst is a bonus when network conditions permit it without infringing on any user's CIR. For this reason, burst frames are marked Discard Eligible (DE) and are queued or discarded when network congestion exists. Although experience has shown that customers can achieve significant burst throughput, it is unreliable and unpredictable and not suitable for real-time applications such as VoIP.

Therefore, the objective is to prevent voice traffic from entering the burst range and being marked DE. One way to accomplish this is to prohibit bursting by shaping the traffic to the CIR and setting the excess burst size (Be—determines the burst range) to zero. However, this also prevents data traffic from using the burst range as well. Another possible alternative is to size the CIR above the peak voice traffic level, and then prioritize the voice traffic so that it is always delivered first. The underlying assumption here is that the network administrator has an expectation of peak voice traffic. By sizing the CIR to meet or exceed the peak voice traffic and then applying priority queuing on the interface so that VoIP is serviced first, we can intuitively assure that voice traffic will not enter the burst range.

The problem with the latter method, however, is that the actual queuing mechanism is not always intuitive. Even though the aggregate voice traffic throughput cannot exceed the CIR, it is possible that a voice packet could be sent in the burst range. The technical workings of this are beyond the scope of this document. But simply stated, it is possible that a voice packet would arrive right after many data packets have already been transmitted in the CIR range, such that the voice packet ends up in the burst range when the router processes it. However, the method is certainly worth trying.

One good piece of knowledge is that most IXCs convert the long-haul delivery of frame relay into ATM. That is, the frame relay PVC is converted to an ATM PVC at the first frame relay switch after leaving the customer's premise. It is not converted back to frame relay until the last frame relay switch before entering the customer's premise. This has significance because ATM has built-in CoS. A customer can contract with a carrier to convert the frame relay PVC into a Constant Bit Rate (CBR) ATM PVC. ATM CBR cells are delivered with lower latency and higher reliability.

As a final note, the reader should understand that under the best circumstances, frame relay is still inherently more susceptible to delay than ATM or TDM. Therefore, after applying the best possible queuing mechanism, one should still expect more delay over frame relay than would be present over ATM or TDM.

Network address translation (NAT)

VoIP does not work well with networks that use NAT because most NAT implementations do not support H.323 protocols. The destination IP address is encapsulated in more than one header: the Q.931, H.225, and IP headers. NAT changes only the address in the IP header, resulting in a mismatch that prohibits the control of calls. Using a firewall to guard against intruders is often suggested, but the firewall should not provide NAT functions for VoIP packets unless it is Q.931 friendly.

Network design recommendations

This section is an amalgam of various converged network vendors' recommendations for best practices in designing a converged IP network.

A network should be designed with the following factors in mind:

- Reliability/redundancy
- Scalability
- Manageability
- Bandwidth.

Supporting voice adds even more considerations when designing a network. These additional considerations include

- Delay
- Jitter
- Loss
- Duplex.

Generally speaking, these concerns dictate a hierarchical network consisting of at most three layers: core,

distribution, and access. Some smaller networks can collapse the functions of several layers into one device.

The core layer is the heart of the network. Its purpose is to forward packets as quickly as possible. It should be designed with high availability in mind. Generally, these high-availability features include redundant devices, redundant power supplies, redundant processors, and redundant links. In the current era, core interconnections increasingly use Gigabit Ethernet.

The distribution layer links the access layer with the core. It is here that QoS features and access lists are applied. Generally, Gigabit Ethernet connects to the core and either Gigabit Ethernet or 100base-TX/FX links connect the access layer. Redundancy is important at this layer, but not as important as in the core.

The access layer connects servers and workstations. Switches at this layer are smaller, usually 24–48 ports. Desktop computers and workstations are usually connected at 10 Mbps (or 100 Mbps) and servers are connected at 100 Mbps (or 1 Gbps). Limited redundancy is used. Some QoS and security features can be implemented here.

For VoIP to work well, WAN links should be properly sized with sufficient bandwidth for voice and data traffic. Each voice call uses between 6.3 kbps and 80 kbps, depending on the desired codec, quality, and header compression used. Interoffice bandwidth demands can be sized using traditional phone metrics such as average call volume, peak volume, and average call length.

Quality of Service also becomes increasingly important with WAN circuits. In this case, QoS can be taken to mean classification and prioritization of voice traffic. Voice traffic should be given absolute priority through the WAN, and if links are not properly sized or queuing strategies are not properly implemented, it will become evident both with the quality and timeliness of voice and data traffic.

There are three technologies that work well with VoIP: ATM, Frame Relay, and point-to-point (PPP) circuits. These technologies all have good throughput, low latency, and low jitter. ATM has the added benefit of enhanced QoS. Frame Relay and PPP links are more economical but lack some of the traffic-shaping features of ATM.

Of the three technologies, Frame Relay is the most difficult WAN circuit to use with VoIP. Congestion in Frame Relay networks can cause frame loss, which can significantly degrade the quality of VoIP conversations. With Frame Relay, proper sizing of the CIR is critical. In a Frame Relay network, any traffic exceeding the CIR is marked DE, and will be discarded at the carrier's option if it experiences congestion in its switches. It is very important that voice packets not be dropped. Therefore, CIR should be sized to average traffic usage. Usually, 25 percent of peak bandwidth is sufficient. Also, SLAs should be established with the carrier that define maximum levels of delay and frame loss, and remediation should the agreed-to levels not be met.

Network management is another important area to consider when implementing VoIP. Because of the stringent requirements imposed by VoIP, it is critical to have an end-to-end view of the network and ways to implement QoS policies globally. Products such as HP Open View Network Node Manager, Visibility, Concord NetHealth, and MRTG will help administrators maintain acceptable service. Should a company not have the resources to implement and maintain network management, outsource companies are springing up to assist with this need.

Common issues

Common network designs that can severely impact network performance when using VoIP include the following:

- Using a flat, non-hierarchical network: This can results in bottlenecks as all traffic must flow across the uplinks (at maximum 1 Gbps) versus traversing switch fabric (up to 256 Gbps). The greater the number of small switches (layers), the greater the number of uplinks and the lower the bandwidth for an individual connection. Under a network of this type, voice performance can quickly degrade to an unacceptable level.
- Multiple subnets on a VLAN: A network of this type can have issues with broadcasts, multicasts, and routing protocol updates. It can greatly impact voice performance and complicate troubleshooting issues.
- Hub-based network: Hubs in a network create some challenges for administrators. The collision domain, the number of ports connected by hubs without a switch or router in between, should be kept as low as possible. The effective (half-duplex) bandwidth available on a shared collision domain is approximately 35 percent of the total bandwidth available.
- Too many access lists: Access lists slow down a router. While they are appropriate for voice networks, care must be taken not to apply them to unnecessary interfaces. A good practice is to model the traffic beforehand, with access lists applied only to the appropriate interface in the appropriate direction, not all interfaces in all directions.

Additional concerns when implementing VoIP include:

- Network Address Translation (NAT): Due to limitations in the H.323 VoIP standard, VoIP conversations rarely work across NAT boundaries. It is important to route voice streams around routers or firewalls running NAT or use a H.323-friendly NAT.

- VPN: The encryption used with VPNs can add significant latency to voice streams, adversely affecting the voice quality. Also, VPNs generally run over the Internet. Because there is no control over QoS parameters for traffic crossing the Internet, voice quality may suffer due to excessive packet loss, delay, and jitter.

In this chapter, we have discussed the special issues involved in the transmission of voice, a real-time application, over an IP network, a non-real-time media. We discussed the needs of the voice packets and how those needs can affect the quality of the voice transmission. We also discussed the various protocols and methodologies used to provide for good-quality voice in an IP network. The last section of the chapter also provided some recommendations for the design and implementation of a converged VoIP network.

The last section of the chapter also provided some recommendations for the design and implementation of a converged VoIP network.

Chapter 3.4

3.4

Channel protection fundamentals

Raouf Hamzaoui, Vladimir Stanković, Zixiang Xiong,
Kannan Ramchandran, Rohit Puri, Abhik Majumdar, and Jim Chou

3.4.1 Introduction

In many ways, the Internet (or a wireless network for that matter) can be regarded simply as a communication channel in a classical communication system. This chapter discusses the fundamentals of channel protection that lie beneath the error control techniques used to communicate multimedia over the Internet and wireless networks.

The goal of a classical communication system is to transfer the data generated by an information source efficiently and reliably over a noisy channel. The basic components of a digital communication system are shown in Figure 3.4-1: a source encoder, channel encoder, modulator, demodulator, channel decoder, and source decoder.

The source encoder removes the redundancy in the digital data produced by the information source and outputs an information sequence. If the information source is analog, its output must be digitized before it is processed by the source encoder. The channel encoder adds redundancy to the information sequence so that channel errors can be detected or corrected. The output of the channel encoder is a finite sequence of symbols called a channel codeword. The set of possible channel codewords is called a channel code. The modulator maps the channel codeword to a signal that is suitable for transmission over a physical channel. The demodulator converts the received signal into a discrete sequence of real numbers of the same length as the channel codeword. In hard decision decoding, each real number in the sequence is mapped to a channel code symbol before being processed by the channel decoder. When the real numbers are left unquantized or quantized to a number of levels that is greater than the size of the channel code alphabet, one speaks of soft decision decoding. The channel decoder tries to recover the input to the channel encoder from the output of the demodulator. Finally, the source decoder produces an estimate of the information sequence.

In this chapter, we look in detail at the basic constituents of a communication system. In Section 3.4.2, we briefly explain the notions of entropy of an information source, mutual information, rate-distortion function, and capacity of a channel. These notions are needed to introduce four fundamental theorems due to Claude Shannon: the noiseless coding theorem [45], the source coding theorem [46], the channel coding theorem [45], and the source-channel coding theorem [46]. Our exposition mainly follows that of McEliece [17], where proofs of the theorems can be found. For simplicity, we focus on discrete memoryless sources and channels. References are provided for extensions and generalizations. Shannon's coding theorems give an insight on what can be achieved by a communication system. Unfortunately, the theorems are not constructive. Optimal codes were shown to exist, but it was not explained how to construct them. The remainder of the chapter is dedicated to practical system design. Section 3.4.3 contains an overview of the most important channel codes. Section 3.4.4 reviews state-of-the-art modulation techniques, focusing on a promising method called hierarchical modulation. Section 3.4.5 considers communication systems where feedback information can be sent from the receiver to the transmitter. In this situation, error control based on error detection and retransmission (automatic repeat request [ARQ]) can be more efficient than error correction alone (forward error correction [FEC]). We review the most important ARQ techniques and discuss hybrid methods that combine ARQ and FEC.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007 Elsevier Inc. All rights reserved

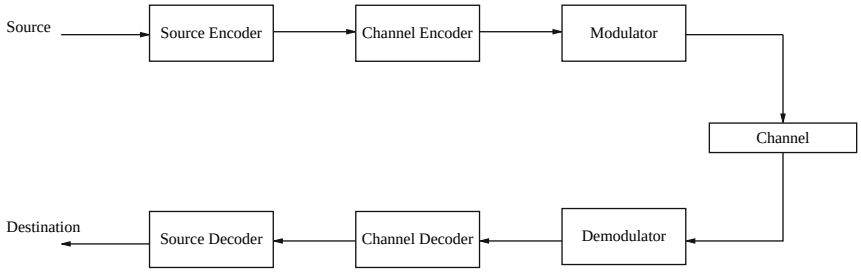


Figure 3.4-1 Basic components of a digital communication system.

3.4.2 Shannon's source and channel theorems

An information source is given by a sequence of random variables X_n, each of which takes values in an alphabet A. We say that the source is *discrete* if the alphabet A is finite or countable. We say that the source is *memoryless* if the random variables X_n are independent and identically distributed. For simplicity, a discrete memoryless source X_n will be denoted by a random variable X whose probability distribution $p(x)$ is the probability distribution common to all the random variables X_n.

Definition 3.4.1 (Entropy of a discrete memoryless source). Let X be a discrete memoryless source with alphabet A. The entropy of X is

$$H(X) = -\sum_x p(x) \log p(x),$$

where the sum is taken over all $x \in A$ for which $p(x) > 0$.

Any base of the logarithm can be used in the expression for the entropy. When this base is 2, the entropy is measured in bits.

The entropy of a discrete memoryless source X measures the amount of uncertainty in the source. Since the entropy is completely defined by the probability distribution of X, we speak also of the entropy of the random variable X.

Given the discrete random variables $X, Y_1, \ldots, Y_n$, the *conditional entropy* $H(X \mid Y_1, \ldots, Y_n)$ is the average uncertainty remaining in X after observing $Y_1, \ldots, Y_n$. Formally, with $H(X \mid y_1, \ldots, y_n) = -\sum_x p(x|y_1, \ldots, y_n) \log p(x|y_1, \ldots, y_n)$ we set

$$H(X \mid Y_1, \ldots, Y_n) = \sum_{y_1, \ldots, y_n} p(y_1, \ldots, y_n) H(X \mid y_1, \ldots, y_n).$$

The difference $I(X;Y) = H(X) - H(X|Y)$ is known as the *mutual information* of the two random variables X and Y. It expresses the amount of information provided by Y on X.

3.4.2.1 Source coding

A code $\mathcal{C}$ of size M over an alphabet B is a set of M words, possibly varying in length, made up of symbols from B. The words in $\mathcal{C}$ are called *codewords*. The length of a codeword $\mathbf{c}$ (i.e., the number of symbols in $\mathbf{c}$) is denoted $|\mathbf{c}|$. When all the codewords in $\mathcal{C}$ have the same length k, we say that $\mathcal{C}$ is a block code of length k.

3.4.2.1.1 Lossless source coding

To compress a discrete source X_n over alphabet A without information loss, one can use a lossless encoding scheme $(\mathcal{C}, f)$, which consists of a code $\mathcal{C}$ (known as a source code) and an injective map $f : A \to \mathcal{C}$ (known as an encoder). A source code $\mathcal{C}$ is *uniquely decodable* if for all positive integers k and all codewords $\mathbf{c}_i$, $\mathbf{d}_i$, $i = 1, \ldots, k$, the equality $\mathbf{c}_1 * \cdots * \mathbf{c}_k = \mathbf{d}_1 * \cdots * \mathbf{d}_k$ implies that $\mathbf{c}_i = \mathbf{d}_i$ for all $i = 1, \ldots, k$. Here $*$ denotes concatenation.

The average codeword length of an encoding scheme $(\mathcal{C}, f)$ with respect to a discrete memoryless source X can be expressed $L = \sum_x p(x)|f(x)|$.

Shannon's noiseless coding theorem [45] says that the entropy of a discrete memoryless source X over alphabet A gives the smallest average number of code symbols (from the alphabet B) needed to losslessly represent one source symbol (from the alphabet A), when the base of the logarithm is $|B|$. Before stating the theorem more precisely, we must define the extension of a discrete source. The kth extension of a source X_n is the source $X_n^k = (X_{(n-1)k+1}, \ldots, X_{nk})$ obtained by blocking X_n into blocks of length k. Thus if X_n is a discrete memoryless source taking values in A, then X_n^k is a discrete memoryless source taking values in A^k, where A^k is the set of all words of length k over A.

Theorem 3.4.1 (Shannon's noiseless coding theorem). *Let X be a discrete memoryless source. Let X^k be its kth extension. Let L_k be the minimum average codeword length over all encoding schemes for X^k whose codes are uniquely decodable. Then*

$$H(X) \le \frac{L_k}{k} \le H(X) + \frac{1}{k}.$$

3.4.2.1.2 Lossy source coding

Suppose now that the symbols generated by a discrete memoryless source X over alphabet A are to be reproduced by symbols from a finite alphabet $\widehat{A}$ called the reproducing alphabet. A single-letter distortion measure d: $A \times \widehat{A} \to \mathbb{R}^+$ measures the distortion $d(x, y)$ when symbol $x \in A$ is reproduced as $y \in \widehat{A}$. The distortion between a word $\mathbf{x} = (x_1, \ldots, x_k) \in A^k$ and a word $\mathbf{y} = (y_1, \ldots, y_k) \in \widehat{A}^k$ is defined as $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{k} d(x_i, y_i)$.

A lossy compression scheme $(\mathcal{C}, k, M, f)$ for the memoryless source X is given by a block code $\mathcal{C}$ of length k and size M over the reproducing alphabet $\widehat{A}$, and a mapping f from A^k to $\mathcal{C}$. This compression scheme allows us to represent any sequence of k source symbols with $[\log_2 M]$ bits. Thus the *rate* of a block code of length k and size M is defined to be $(\log_2 M)/k$ bits per symbol.

Example 1. Let $A = \{0, 1\}$, $\widehat{A} = \{0, 1\}$, and $\mathcal{C} = \{00, 11\}$. Then the mapping $f : A^2 \to \mathcal{C}$ given by $f(00) = 00$, $f(01) = 00$, $f(10) = 11$, $f(11) = 11$ defines a lossy compression scheme $(\mathcal{C}, 2, 2, f)$ for the source X. By using the binary representation $00 \mapsto 0$ and $11 \mapsto 1$, any sequence of two source symbols can be represented by one bit.

The average distortion of the lossy compression scheme $(\mathcal{C}, k, M, f)$ with respect to the source X and the single-letter distortion measure d is

$$D(\mathcal{C}, k, M, f) = \frac{1}{k} \sum_{\mathbf{x} \in A^k} p(\mathbf{x}) d(\mathbf{x}, f(\mathbf{x})),$$

where for $\mathbf{x} = (x_1, \ldots, x_k), p(\mathbf{x}) = \prod_{i=1}^{k} p(x_i)$.

Example 2. In Example 1, suppose that the distortion measure d is given by $d(x, y) = 0$ if $x = y$ and $d(x, y) = 1$, otherwise. Suppose also that $Pr\{X = 0\} = p(0) = p$, where $0 < p < 1/2$. Then the average distortion of the lossy compression scheme $(\mathcal{C}, 2, 2, f)$ is $p(1 - p)$.

Let Y be a random variable that is jointly distributed with X according to the joint probability distribution $p(x, y) = p(x)p(y|x)$ for some conditional probability distribution $p(y|x) = p_{Y|X}(y|x)$. As a function of the conditional probability distribution $p_{Y|X}$, the average distortion when X is reproduced as Y is thus

$$D(p_{Y|X}) = \sum_{x \in A, y \in \widehat{A}} p(x) p(y|x) d(x, y).$$

Note that the smallest value that can be taken by $D(p_{Y|X})$ is

$$D_{\min} = \sum_{x \in A} p(x) \min_{y \in \widehat{A}} d(x, y).$$

Definition 3.4.2 (Rate–distortion function). Let X be a discrete memoryless source over alphabet A, and let $\widehat{A}$ be a reproducing alphabet. The rate–distortion function of the source X with respect to a single-letter distortion measure $d : A \times \widehat{A} \to \mathbb{R}^+$ is the function

$$R(D) = \min_{p_{Y|X} : D(p_{Y|X}) \le D} I(X; Y),$$

for all $D \ge D_{\min}$, where the minimum is taken over all conditional distributions $p(y|x)$ subject to the constraint $D(p_{Y|X}) \le D$.

Henceforth we assume that the base of the logarithm in the mutual information is 2, so that $R(D)$ is measured in bits.

We can now state Shannon's source coding theorem [46].

Theorem 3.4.2 (Shannon's source coding theorem). *Let X be a discrete memoryless source over alphabet A, and let $\widehat{A}$ be a reproducing alphabet. Let $R(D)$ be the rate-distortion function of the source X with respect to a single-letter distortion measure $d : A \times \widehat{A} \to \mathbb{R}^+$. Then for any $D' > D$ and $R' > R(D)$, there exists a lossy compression scheme (C, k, M, f) such that $M \le 2^{\lfloor kR' \rfloor}$ and $D(C, k, M, f) < D'$.*

The theorem roughly says that one can reproduce the source symbols with an average distortion that is smaller than D by spending no more than $R(D)$ bits per source symbol.

3.4.2.2 Channel coding

A discrete channel takes at every time unit n a symbol from a finite input alphabet A_X and outputs symbols from a finite output alphabet A_Y.

Let X_n be the random variable that gives the nth input symbol of the channel and let Y_n be the random variable that gives the nth output symbol of the channel. We say that the channel is memoryless if $Pr\{Y_n = y \mid X_n = x\} = p(y|x)$ is independent of n for all $y \in A_Y$ and $x \in A_X$, the transition probabilities $p(y|x)$ satisfy $p(y|x) \ge 0$ and $\sum_{y \in A_Y} p(y|x) = 1$, and $Pr\{Y_1 = y_1 \ldots, Y_n = y_n \mid X_1 = x_1, \ldots, X_n = x_n\} = \prod_{i=1}^{n} p(y_i|x_i)$ for all n.

Definition 3.4.3 (Channel capacity). The capacity of a discrete memoryless channel with input alphabet A_X, output alphabet A_Y, and transition probabilities $p(y|x)$, $y \in A_Y$, $x \in A_X$ is

$$C = \max_{p_X} I(X; Y),$$

where X is a random variable that gives the input symbol to the channel and Y is a random variable that gives the corresponding output symbol according to the joint probability distribution $p(x, y) = p(x)p(y|x)$, for some source distribution $p(x) = p_X(x)$.

Since we assume that the base of the logarithm in the mutual information is 2, C is measured in bits.

Example 3. The binary symmetric channel (BSC) is a discrete memoryless channel with $A_X = A_Y = \{0,1\}$ and transition probabilities $p(0|1) = p(1|0) = p$ and $p(0|0) = p(1|1) = 1 - p$. When a bit is sent over the BSC, it is either corrupted with probability p or correctly received with probability $1 - p$. It is easy to prove that the capacity of the BSC is $1 + p \log p + (1 - p) \log(1 - p)$ [15].

Example 4. The binary erasure channel (BEC) is a discrete memoryless channel with $A_X = \{0, 1\}$, $A_Y = \{0, 1, ?\}$ and transition probabilities $p(?|0) = p(?|1) = p$ and $p(0|0) = p(1|1) = 1 - p$. When a bit is sent over the BEC, it is either erased with probability p or correctly received with probability $1 - p$. The capacity of the BEC is $1 - p$ [15].

We now consider the situation when the symbols generated by a discrete source U with source alphabet A are to be sent over a discrete channel with input alphabet A_X and output alphabet A_Y. To protect the source symbols against transmission errors, redundancy is added. For this, we use a channel code, which is a block code C of length n over A_X. We also use a channel encoding scheme, which is an injective function from A^k to C. Using the channel encoding scheme, blocks $(u_1, \ldots, u_k)$ of source symbols of length k are mapped to channel codewords of length n. The channel codewords are then sent over the channel where errors may occur. Next, a function $\mathbf{g}$ from A_Y^n to C called a channel decoding scheme, is used to map a received word $\mathbf{y} = (y_1, \ldots, y_n)$ to a channel codeword. Finally, this channel codeword is mapped to the corresponding source word $(\widehat{u}_1, \ldots, \widehat{u}_k)$ (since the encoding scheme is injective, this source word is unique when it exists). The rate of transmission of this system (or code rate of the code C) is $\frac{k}{n}$. It characterizes the speed with which source information is transmitted over the channel or equivalently the redundancy introduced by the channel code.

An ideal channel decoding scheme for this system minimizes the probability of a word decoding error

$$P_e = \sum_{\mathbf{c} \in \mathcal{C}} \sum_{\mathbf{y} \in A_Y^n : g(\mathbf{y}) \neq \mathbf{c}} Pr\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{c}\} Pr\{\mathbf{X} = \mathbf{c}\},$$

where $\mathbf{Y} = (Y_1, \ldots, Y_n)$ and $\mathbf{X} = (X_1, \ldots, X_n)$. This is realized with *maximum a posteriori decoding*, where the received word $\mathbf{y} = (y_1, \ldots, y_n)$ is mapped to a channel codeword $\mathbf{c} = (c_1, \ldots, c_n)$ that maximizes the probability $\boldsymbol{Pr}\{\mathbf{X} = \mathbf{c} | \mathbf{Y} = \mathbf{y}\}$. In practice, however, one uses *maximum-likelihood decoding*, where $\mathbf{y}$ is mapped to a channel codeword $\mathbf{c}$ that maximizes the probability $\boldsymbol{Pr}\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{c}\}$. It is easy to see that maximum-likelihood decoding is equivalent to maximum *a posteriori* decoding when all channel codewords are generated with the same probability. Another important decoding scheme is known as *minimum distance decoding* where the received word $\mathbf{y} = (y_1, \ldots, y_n)$ is mapped to a channel codeword $\mathbf{c} = (c_1, \ldots, c_n)$ that has smallest Hamming distance to $\mathbf{y}$. Here the Hamming distance is defined as follows.

Definition 3.4.4 (Hamming distance). Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two words of the same length. The Hamming distance $d_H(\mathbf{x}, \mathbf{y})$ between $\mathbf{x}$ and $\mathbf{y}$ is equal to the number of indices $k \in \{1, 2, \ldots, n\}$ such that $x_k \neq y_k$.

Given a BSC with bit error probability p, we have

$$Pr\{\mathbf{Y} = \mathbf{y} | \mathbf{X} = \mathbf{c}\} = p^{d_H(\mathbf{c}, \mathbf{y})} (1 - p)^{n - d_H(\mathbf{c}, \mathbf{y})}. \tag{3.4.1}$$

If $0 < p < 1/2$, the probability in (3.4.1) is largest when $d_H(\mathbf{c}, \mathbf{y})$ is smallest. Therefore minimum distance decoding and maximum-likelihood decoding are equivalent for this channel.

An alternative to minimizing the probability of a word decoding error is to minimize the information symbol error rate p_e defined as $p_e = \frac{1}{k} \sum_{i=1}^{k} p_e^{(i)}$, where $p_e^{(i)} = Pr\{\widehat{u}_i \neq u_i\}$. When the symbols are bits, the information symbol error rate is called the information *bit error rate* (BER). Note that $\frac{1}{k} P_e \leq p_e \leq P_e$. To minimize the information symbol error rate, the decoder uses the *symbol maximum a posteriori* (MAP) rule, where for $i = 1, \ldots, k$, the reconstructed information symbol $\widehat{u}_i$ is computed as a symbol $u \in A$ that maximizes the *a posteriori* probability $Pr\{U_i = u \mid \mathbf{Y} = \boldsymbol{y}\}$. Here $\boldsymbol{U}_i$ is the random vector that corresponds to the information symbol $\mathrm{u_i}$, $i = 1, \ldots, k$.

The channel coding theorem [45] states that the source information can be transmitted reliably over a noisy channel, provided the rate of transmission is below the capacity of the channel. In other words, any rate below the channel capacity is achievable.

Theorem 3.4.3 (Shannon's channel coding theorem). *Consider a discrete memoryless channel with input alphabet A_X and capacity C. For any positive number $R < C$ and $\varepsilon > 0$, there exists a channel code $\mathrm{C} = \{\mathbf{c}_1, \ldots, \mathbf{c}_M\}$ of length n over A_X and a channel decoding scheme g such that*

1) $M \geq 2^{\lfloor Rn \rfloor}$

2) *If codeword $\mathbf{c}_i$ is sent over the channel and word $\mathbf{y}$ is received, then $Pr\{g(\mathbf{y}) \neq \mathbf{c}_i\} < \varepsilon$ for all $i = 1, \ldots, M$.*

3.4.2.3 Source-channel coding

Suppose now that the output of the channel decoding scheme is mapped to a word of length k over a reproduction alphabet $\widehat{A}$. The average distortion of the resulting transmission system is $\frac{1}{k} E[d(\mathbf{U}, \mathbf{V})]$, where the random vector $\mathbf{U}$ describes a word of k successive

source symbols, the random vector $\mathbf{V}$ describes the corresponding word of k reconstructed symbols, and E denotes the expectation operator.

The source-channel coding theorem [46] says what a system can achieve in terms of average distortion and rate of transmission.

Theorem 3.4.4 (Shannon's source-channel coding theorem). *Given a discrete memoryless source characterized by rate–distortion function $R(D)$, a discrete memoryless channel characterized by capacity $C > 0$, any $D > D_{\min}$, and any $r < C/R(D)$, there exist for sufficiently large k and n an encoding scheme that maps source words of length k into channel words of length n and a decoder that maps channel output words of length n into reproduced words of length k such that the expected distortion is at most D and the transmission rate k/n is at least r.*

The encoding scheme promised by the theorem is a concatenation of a lossy compression scheme and a channel encoding scheme. The theorem is also known as the separation theorem because the lossy compression scheme and the channel encoding scheme can be designed independently.

3.4.2.4 Extensions

Shannon's theorems can be extended to more general information sources. For example, we say that a discrete source X_n is *stationary* if the random process X_n is stationary. The nth marginal entropy of a stationary source is $H_n = H(X_n|X_{n-1,\ldots,}X_1)$. One can show that when the source is stationary, the sequence H_n is decreasing and bounded below by zero. This allows us to define the entropy of a stationary source as follows.

Definition 3.4.5 (Entropy of a stationary source). Let X_n be a stationary source. The entropy of the source (also often called the entropy rate of the source) is defined as $\overline{H} = \lim_{n\to\infty} H_n$.

With this definition, Shannon's noiseless coding theorem can be extended to stationary sources that satisfy the asymptotic equipartition property [3]. The source coding theorem can also be extended to sources with abstract alphabets, including the set of real numbers in particular [22].

Shannon's channel coding theorem can be extended to other channels, the most famous one being the *additive white Gaussian noise* (AWGN) channel. In the time-discrete AWGN channel, both the channel input alphabet A_X and the channel output alphabet A_Y are the set of real numbers $\mathbb{R}$. The relationship between the random variable X_n that gives the nth input to the channel and the random variable Y_n that gives the nth output of the channel is given by $Y_n = X_n + Z_n$, where $\{Z_n\}$ is a sequence of independent, identically distributed, Gaussian random variables with mean 0 and variance $N_0/2$. One can show [15] that for this channel reliable transmission is possible as long as the rate of transmission is smaller than the capacity

$$C = \frac{1}{2}\log_2\left(1 + \frac{2P}{N_0}\right) \text{ bits per transmission,}$$

where P is a constraint on the expected value of the random variable X_n^2. If we denote by R the rate of transmission, by $E_s = P$ the symbol energy, and by $E_b = E_s/R$ the energy per bit, then the condition $R < C$ gives $E_b/N_0 > \frac{2^{2R}-1}{2R}$ for reliable transmission. Here E_b/N_0 is called the bit energy to noise spectral density ratio and $\frac{2^{2R}-1}{2R}$ is the Shannon bound. Since $R > 0$, we must also have $E_b/N_0 > \log_e 2$ or $10\log_{10} E_b/N_0 > -1.6$ dB, which is called the theoretical Shannon limit.

In the time-continuous AWGN channel, the relationship between the transmitted signal $s(t)$ (the output of the modulator) and the received signal $r(t)$ (the input of the demodulator) is $r(t) = s(t) + n(t)$, where $n(t)$ is a white Gaussian noise. The capacity of a band-limited AWGN channel is [15]

$$C = W\log_2\left(1 + \frac{P}{N_0 W}\right) \text{ bits per second,}$$

where W is the channel bandwidth in Hz, $N_0/2$ is the power spectral density of the noise, and P is a constraint on the average power. When P is much smaller than N_0W, the channel is called a wideband AWGN channel. One can prove [11] that if binary modulation is used and the demodulated signal is sampled at a rate of $2W$, then E_b/N_0 must be larger than the practical Shannon limit of 0.2 dB to achieve a BER of 10^{-5} for a rate of transmission $R = 1/2$.

3.4.3 Channel coding and error control for bit errors and packet losses

Channel codes can be divided into two classes: linear and nonlinear. Linear codes are easier to implement and, as a result, have received a greater amount of attention historically. We will also confine our attention to linear codes in this section. We first describe linear block codes, including cyclic redundancy check (CRC) codes for error detection, Reed–Solomon codes, low–density parity–check (LDPC) codes, irregular repeat-accumulate (IRA) codes, tornado codes, digital fountain codes, and lattice codes. We then describe convolutional codes, rate-compatible punctured convolutional (RCPC) codes, and turbo codes. We discuss the properties of these codes and mention efficient algorithms for encoding and decoding,

emphasizing their computational complexity. We also explain how the problem of burst errors can be alleviated with interleaving.

3.4.3.1 Linear block codes

In a linear block code, the codeword symbols are taken from a field. A formal definition of a field is beyond the scope of this text and may be found in a mathematics book on abstract algebra (see [12]). Informally, a field consists of a set of elements, together with two operations called addition and multiplication that must fulfill a given number of properties. Some examples of well-known fields are the set of real numbers and the set of rational numbers. These fields are known as infinite fields because they contain an infinite number of elements. Linear block codes, however, typically consist of elements from a finite field. In particular, consider the finite field $GF(2) = \{0, 1\}$. The addition operation for $GF(2)$ is modulo-2 addition and the multiplication operation is defined similarly to the multiplication operation for two binary numbers:

$$0+0=0 \quad 0+1=1 \quad 1+1=0$$
$$0*0=0 \quad 0*1=0 \quad 1*1=1.$$

Finite fields are also called Galois fields. The size of a Galois field must be a power of a prime. Conversely, for any prime power q, one can construct a Galois field of size q. Let $GF(q)$ be a finite field of size q and let n be a positive integer. Then it is easy to check that $[GF(q)]^n$ is a linear space over $GF(q)$. An (n, k) linear block code $\mathcal{C}$ over $GF(q)$ is a k-dimensional linear subspace of the linear space $[GF(q)]^n$. In particular, for any two codewords $\mathbf{c}_1$, $\mathbf{c}_2 \in \mathcal{C}$, the sum of the codewords is also a codeword, $\mathbf{c}_1 + \mathbf{c}_2 \in \mathcal{C}$. Since $\mathcal{C}$ is a k-dimensional linear space, we can find a set of k basis vectors so that every codeword can be expressed as a linear combination of the basis vectors. In vector-matrix notation, we can express every codeword $\mathbf{c}$ of $\mathcal{C}$ as

$$\mathbf{c} = \mathbf{uG}, \tag{3.4.2}$$

where $\mathbf{u}$ is a $1 \times k$ vector of field elements and $\mathbf{G}$ is a $k \times n$ matrix whose k rows are k basis vectors. The matrix $\mathbf{G}$ is known as a generator matrix and elementary row operations can be performed on $\mathbf{G}$ to form another matrix $\mathbf{G}'$ that will generate an equivalent code. If $\mathbf{G}$ is manipulated to be of the form $\mathbf{G} = [\mathbf{I}_k|\mathbf{P}]$ where $\mathbf{I}_k$ is the $k \times k$ identity matrix and $\mathbf{P}$ is a $k \times (n-k)$ matrix, then $\mathbf{G}$ is said to be in systematic form and the first k symbols of the codeword $\mathbf{c}$ will be identical to the k symbols of $\mathbf{u}$. The final $n-k$ symbols of $\mathbf{c}$ are referred to as parity symbols.

The performance of a block code is often measured by the number of errors that it can correct or the amount of noise that it can remove. The performance is usually dependent on two things: (1) the decoder that is used to decode a received word to a codeword and (2) the distance between each pair of codewords in the block code. Let us first consider the distance between a pair of codewords. For block codes, the Hamming metric or the Euclidean metric is usually used to measure the distance between pairs of codewords. The Hamming distance is useful for measuring the distance between two codewords whose symbols belong to a finite field. Sometimes in communications applications, however, each field element of the codeword is mapped to a real number. In such scenarios, it is useful to use a Euclidean metric to determine the distance between codewords. For example, if each codeword, $\mathbf{c}_i = (c_{i,1}, c_{i,2}, \ldots, c_{i,n})$, is mapped to a vector of real numbers, $\mathbf{r}_i = (r_{i,1}, r_{i,2}, \ldots, r_{i,n})$, then the Euclidean distance between two codewords $\mathbf{c}_i$, $\mathbf{c}_j$ may be defined as

$$d_E(\mathbf{c}_i, \mathbf{c}_j) = \left[(r_{i,1} - r_{j,1})^2 + (r_{i,2} - r_{j,2})^2 + \cdots + (r_{i,n} - r_{j,n})^2\right]^{1/2} \tag{3.4.3}$$

Now, if we let $d_{\min}$ represent the minimum distance between any pair of codewords, and if an arbitrary codeword is transmitted over a noisy channel, then the codeword may be successfully recovered if the decoder decodes the received word to the closest codeword and the amount of noise is less than $d_{\min}/2$. Note that if the block code is linear, then the minimum distance $d_{\min}$ is simply the smallest weight of a nonzero codeword. Here the weight of a codeword is the number of its nonzero symbols.

For the Hamming distance, successful decoding translates into there being less than $d_{\min}/2$ changes to the symbols of the original codeword, where $d_{\min}$ represents the minimum Hamming distance between any pair of codewords. For the Euclidean distance, successful decoding translates into the magnitude of the noise being less than $d_{\min}/2$. For the aforementioned, we may visualize the correct decoding region of each codeword to be a sphere with radius $d_{\min}/2$ (as in Figure 3.4-2), and thus if a codeword is corrupted by noise, as long as the noise does not perturb the codeword to be outside of its correct decoding region, then successful decoding will be guaranteed. It is apparent that for a given n and k, it is desirable to find a code that maximizes $d_{\min}$.

Once a code with parameters n, k, and $d_{\min}$ is found, efficient encoding and decoding algorithms for generating the code are necessary to enable the code to be practical.

In general, if the minimum distance of a linear code is t, then the receiver can detect up to $t - 1$ transmission errors. However, by using minimum distance decoding, where the received word is decoded to a nearest codeword, a linear code of minimum distance t allows the correction of up to $\lfloor \frac{t-1}{2} \rfloor$ errors. It can be shown that

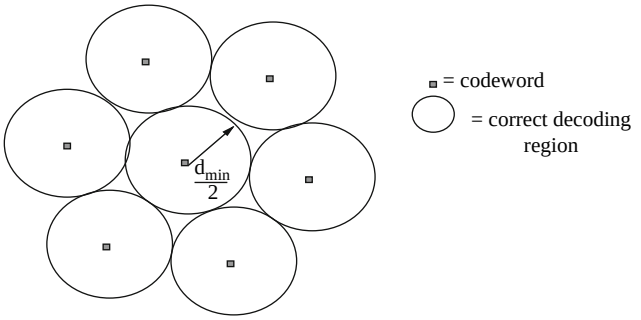


Figure 3.4-2 Example of a codebook that consists of several codewords with a minimum distance of d_{min}. The correct decoding regions are shown as spheres centered around the codewords with a radius of $d_{min}/2$.

a linear code of minimum distance t can simultaneously correct e_c errors and detect e_d errors if $e_c + e_d \leq t - 1$ with $e_c \leq e_d$. Moreover, it can correct e_e erasures and e_c errors simultaneously if $e_e + 2e_c \leq t - 1$.

The minimum distance of an (n, k) linear code must be less than or equal to $n - k + 1$. Linear (n, k) codes whose minimum distance is equal to $n - k + 1$ are called maximum-distance separable (MDS) codes.

Linear codes can be simply modified to obtain new linear codes. Puncturing a linear code consists of removing a number of coordinate positions from each codeword. If an (n, k) linear MDS code is punctured, then the resulting code is an $(n - 1, k)$ linear MDS code. Shortening a linear code consists of keeping only codewords with the same symbol in a given position and then deleting this position. If an (n, k) linear MDS code is shortened by keeping only codewords with the zero symbol in a given position, then the resulting code is an $(n - 1, k - 1)$ linear MDS code.

The encoding scheme for an (n, k) linear code can be implemented in $O(n^2)$ time. However, there is no efficient way to decode a general linear code with maximum-likelihood decoding [8]. Usually one uses syndrome decoding. To explain syndrome decoding, we must introduce the parity check matrix. The parity check matrix $\mathbf{H}$ of an (n, k) linear code with generator matrix $\mathbf{G}$ is an $(n - k) \times n$ matrix whose rows are orthogonal to the rows of the generator matrix, that is,

$$\mathbf{G}\mathbf{H}^T = \mathbf{0}. \tag{3.4.4}$$

The parity check matrix may be viewed as a generator matrix for a code that lies in the null space of $\mathbf{G}$. It is clear that for any codeword $\mathbf{c}$ that is generated by $\mathbf{G}$,

$$\mathbf{c}\mathbf{H}^T = \mathbf{0}. \tag{3.4.5}$$

Now, if we add an error vector $\mathbf{e}$ to $\mathbf{c}$, then

$$(\mathbf{c} + \mathbf{e})\mathbf{H}^T = \mathbf{0} + \mathbf{e}\mathbf{H}^T = \mathbf{s}, \tag{3.4.6}$$

where we call $\mathbf{s}$ the syndrome of $(\mathbf{c} + \mathbf{e})$. If we let each syndrome correspond to an error vector, then the function of a syndrome decoder is to first compute the syndrome of the received vector and then subtract the corresponding error vector from the received vector. Another way of viewing syndrome decoding is through a standard array [36]. A standard array of a binary code is formed by setting aside a $2^{n-k} \times 2^k$ array and populating the first row of the array with all 2^k possible codewords with the all-zero codeword occupying the first column of the first row. Next, we generate all possible weight 1 error vectors and add each error vector to the first row to generate another row. This process is continued by increasing the weight of the error vector and filling the rows until the entire array is populated as in Table 3.4-1. If the result of an addition of an error vector with the first row equals a row that is already in the standard array, then the error vector is skipped and the next error vector is used to generate further rows. The result will be an array, where each row corresponds a shift of all of the codewords by an error vector. The first column will contain the error vectors and each row may be indexed by the syndrome. Therefore, syndrome decoding may be viewed as indexing a row of the standard array and then adding the first element of the row to the received vector.

Example 5. Consider a (3,1) binary repetition code. This block code consists of two codewords, {000,111}. A generator matrix for the code is

$$\mathbf{G} = [1\ 1\ 1]. \tag{3.4.7}$$

The corresponding parity check matrix is

$$\mathbf{H} = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{bmatrix}. \tag{3.4.8}$$

And we can tabulate the standard array as follows:

Table 3.4-1 Standard array of a binary code. The first row contains all codewords of the code. Each following row is formed by taking a minimum weight vector, adding it to the first row, and then checking if the resulting addition is already part of the standard array. If the resulting addition is not part of the standard array, then it is added as a new row to the standard array. This process is continued until the standard array is filled.

$\mathbf{c}_1 = \mathbf{0}$	$\mathbf{c}_2$	…	$\mathbf{c}_{2^k}$
$\mathbf{e}_1$	$\mathbf{e}_1 + \mathbf{c}_2$ …	…	$\mathbf{e}_1 + \mathbf{c}_{2^k}$
$\mathbf{e}_2$	$\mathbf{e}_2 + \mathbf{c}_2$ …	…	$\mathbf{e}_2 + \mathbf{c}_{2^k}$
$\mathbf{e}_3$	$\mathbf{e}_3 + \mathbf{c}_2$ …	…	$\mathbf{e}_3 + \mathbf{c}_{2^k}$
⋮	⋮	⋮	⋮
$\mathbf{e}_{2^{n-k}-1}$	$\mathbf{e}_{2^{n-k}-1}+\mathbf{c}_2$ …	…	$\mathbf{e}_{2^{n-k}-1}+\mathbf{c}_{2^k}$

$$\left\{\begin{array}{ll}000 & 111\\ 001 & 110\\ 010 & 101\\ 100 & 011\end{array}\right\}. \tag{3.4.9}$$

Notice that the minimum weight codeword is (1,1,1) and therefore the minimum distance of the code is 3. This implies that one error may be *corrected* if a minimum distance decoder is used for decoding. Alternatively, up to two errors may be *detected*.

Since the decoding of general linear block codes is not efficient, special classes of linear codes with fast decoding algorithms were developed. The most popular of these is the class of cyclic codes. An (n, k) linear code is a cyclic code if for each codeword $(c_1,\ldots,c_{n-1}, c_n)$, the right shift $(c_n,c_1,\ldots,c_{n-1})$ is also a codeword.

3.4.3.1.1 CRC codes

CRC codes are shortened cyclic binary codes used for error detection. Given a generator polynomial $g(x) = \sum_{i=0}^{r} g_i x^i$, $g_i \in \{0, 1\}$ of degree r, the codeword for a binary information sequence $\mathbf{u} = (u_1,\ldots,u_k)$ is the concatenation $\mathbf{u} * \mathbf{p}$ of $\mathbf{u}$ and the word $\mathbf{p}$ of length r associated to the polynomial $p(x) = x^r u(x) \bmod g(x)$. Here we use the unique correspondence between a word $\mathbf{w} = (u_1, \ldots, u_m)$ of length m and the polynomial $w(x) = w_1 + w_2 x + \ldots + w_m x^{m-1}$ of degree at most $m - 1$. Suppose that the codeword $\mathbf{u} * \mathbf{p}$ is sent over a binary symmetric channel and let $\mathbf{u}' * \mathbf{p}'$ be the received word. Here $\mathbf{u}'$ and $\mathbf{p}'$ are words having the same length as $\mathbf{u}$ and $\mathbf{p}$, respectively. Then the decoder computes $p''(x) = x^r u'(x) \bmod g(x)$ and declares an error if $p''(x)$ is not equal to $p'(x)$. Some of the most popular generator polynomials are the CRC-12 polynomial $1 + x + x^2 + x^3 + x^{11} + x^{12}$, the CRC-16 polynomial $1 + x^2 + x^{15} + x^{16}$, and the CRC-CCITT polynomial $1 + x^5 + x^{12} + x^{16}$. A CRC code with generator polynomial $g(x) = \sum_{i=0}^{r} g_i x^i, g_0 \neq 0, g_r \neq 0$ can detect any burst error of length $k \leq r$. Agarwal and Ivanov [2] provided an $O(nm2^{r+m})$ algorithm for computing the probability of undetected error for a CRC code of length n whose generator polynomial has degree r and m nonzero coefficients. The encoding and decoding of CRC codes can be efficiently implemented with shift register circuits.

3.4.3.1.2 Reed–Solomon codes

Reed–Solomon codes are nonbinary linear block codes over a finite field $GF(q)$. Let α be an element of order n in $GF(q)$ [i.e., n is the smallest positive integer such that $\alpha^n = 1$, where 1 is the identity element for the multiplication in $GF(q)$]. Let $r \in \{1, \ldots, n\}$. The set of all vectors $(c_0, \ldots, c_{n-1})$ in $[GF(q)]^n$ such that $\sum_{i=0}^{n-1} c_i \alpha^{ij} = 0$, $j = 1, \ldots, r$, is called a Reed–Solomon code of redundancy r over $GF(q)$. This code is an $(n, n - r)$ cyclic code of minimum distance $r + 1$. Thus, the Reed–Solomon codes are MDS codes. Therefore, an (n, k) Reed–Solomon code can correct e_0 symbol erasures and e_1 symbol errors simultaneously if $e_0 + 2e_1 \leq n - k$. In particular, in a channel where only erasures can occur, all codeword symbols of an (n, k) Reed–Solomon codeword can be correctly recovered if at least k symbols are received.

Reed–Solomon codes are suitable for the correction of burst errors. An (n, k) Reed–Solomon code can be decoded in $O(n^2)$ time with Berlekamp's iterative algorithm [7]. Guruswami and Sudan [23] developed a polynomial-time algorithm for Reed–Solomon codes that finds a list of all codewords within a distance $\lceil n - \sqrt{n(k-1)} - 1 \rceil$ from a received word. Thus, the algorithm is guaranteed to determine the list of all potentially sent codewords if at most $\lceil n - \sqrt{n(k-1)} - 1 \rceil$ errors occurred during transmission. The complexity of the algorithm is $O(n^{15})$ if exactly $\lceil n - \sqrt{n(k-1)} - 1 \rceil$ errors occurred and only $O(n^3)$ otherwise. The algorithm of Berlekamp [7] is a hard-decision decoding algorithm, which does not exploit all available information at the receiver when the demodulator allows soft decisions. Efficient soft-decision decoding algorithms for Reed–Solomon codes were proposed by Koetter and Vardy [28] and Jiang and Narayanan [26]. For example, the algorithm of Jiang and Narayanan [26] outperforms hard-decision decoding by up to 3.1 dB at decoding error probability 10^{-5} when decoding a (15, 7) Reed–Solomon code over a binary-input AWGN channel.

3.4.3.1.3 LDPC codes

LDPC codes were introduced by Gallager [21]. They have attracted increased interest since MacKay and Neal [34,35] reported their outstanding performance on a binary-input AWGN channel. An (n, k) LDPC code is a linear code with a sparse parity-check matrix $H = (h_{ij})$. It can also be described with a bipartite graph, called Tanner graph, whose set of nodes consists of variable nodes and check nodes. Variable nodes correspond to the n codeword symbols, while check nodes correspond to the $(n - k)$ equations defined by the parity-check constraint. A variable node is connected to a check node if the codeword symbol corresponding to the variable node is involved in the parity equation defining the check node. That is, check node i is connected to variable node j if $h_{ij} = 1$. In a regular LDPC code, each column has the same number d_v of 1's and each row has the same number d_c of 1's. Thus, in the Tanner graph of the code each variable node has degree d_v and each check code has degree d_c, as shown in Figure 3.4-3. In an irregular LDPC code, the degrees of the variable nodes and check nodes are chosen according to some nonuniform distribution. Efficient encoding of LDPC codes is discussed in [40],

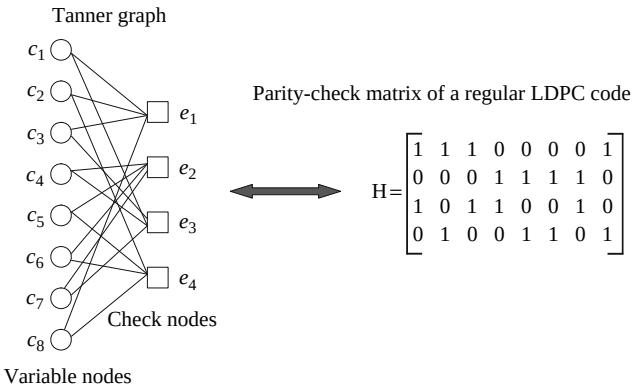


Figure 3.4-3 Tanner graph of a regular LDPC code of length 8. The degree of the variable nodes is $d_v = 2$, and the degree of the check nodes is $d_c = 4$.

where it is shown in particular that some of the best LDPC codes can be encoded in $O(n)$ time with high probability. Because of the sparseness of their Tanner graph, LDPC codes can be decoded in $O(n)$ time with an iterative procedure known as probabilistic decoding [21], message passing, sum–product algorithm [49], or belief propagation [37]. These algorithms alternately pass information between adjacent variable nodes and check nodes to compute estimates of the *a posteriori* probabilities of the codeword symbols. The decoded codeword is based on the estimates obtained after convergence or if a maximum number of iterations is reached. Chung *et al.* [14] were able to design a rate-$\frac{1}{2}$ irregular LDPC code of length 10^7 bits that is only 0.04 dB away from the Shannon limit for a binary-input AWGN channel and a bit error rate of 10^{-6}.

3.4.3.1.4 IRA codes

Irregular repeat–accumulate (IRA) codes were introduced by Jin and colleagues [27] as a generalization of the repeat–accumulate (RA) codes of [16]. IRA codes can be encoded in linear time. They are decoded with the sum–product algorithm, achieving on the binary-input AWGN channel a performance competitive with that of the best LDPC codes of comparable complexity.

3.4.3.1.5 Tornado codes

Tornado codes [31,32] are (n, k) erasure codes that allow encoding and decoding with time complexity linear in the block length n. This speed-up over Reed–Solomon codes is obtained at the cost that slightly more than k encoding symbols are required to reconstruct all k information symbols. More precisely, Luby *et al.* [32] prove that for any $\varepsilon > 0$, one can construct a Tornado code that recovers all k information symbols from only $(1 + \varepsilon)k$ encoding symbols with probability $1 - O(n^{-3/4})$.

3.4.3.1.6 Digital fountain codes

Luby [30] recently introduced a new class of powerful erasure correcting codes called Luby Transform (LT) codes. LT codes are rateless in the sense that a potentially limitless stream of encoding symbols (or a digital fountain) can be generated for a given information sequence. Thus, in contrast to classical block codes, one need not design the code *a priori* for a fixed n. With LT codes, each encoding symbol can be generated from k information symbols in $O(\log k)$ time on average, and one can recover all k information symbols from $k + O(\sqrt{k}\log^2(k/\delta))$ encoding symbols with probability $1 - \delta$ in $O(k \log k)$ time, on average.

By concatenating an LDPC code as an outer code and an LT code as an inner code, Shokrollahi [47] was able to construct rateless codes called Raptor codes whose erasure correcting performance is similar to that of LT codes, but can be encoded and decoded in only $O(k)$ time.

3.4.3.1.7 Lattice codes

Codes over finite fields can also be interpreted as codes over real numbers by mapping each element of the finite field to a real number. For example, in the binary (3, 1) repetition code, the binary digit 0 can be mapped to the real value $-a$ and the binary digit 1 can be mapped to the real value $+a$ so that the two codewords are $(-a, -a, -a)$ and $(+a, +a, +a)$. Minimum distance decoding then means decoding the received vector to the codeword that is closest in Euclidean distance. In this section, we will consider a class of codes called lattice codes, which contain codewords that are amenable to Euclidean distance decoding instead of Hamming distance decoding.

Informally, a lattice Λ is an infinite regular array of points that covers an m-dimensional space uniformly. A lattice is defined by a set of basis vectors so that any point in the lattice can be represented as a linear combination of the basis vectors. More precisely, if the basis vectors are given as $\mathbf{v}_1 = (v_{1,1}, v_{1,2}, \ldots, v_{1,m})$, $\mathbf{v}_2 = (v_{2,1}, v_{2,2} \ldots, v_{2,m})$, $\ldots$, $\mathbf{v}_n = (v_{n,1}, v_{n,2}, \ldots, v_{n,m})$ where $m \geq n$, then we can define a generator matrix, $\mathbf{G}$, to be a matrix that contains the basis vectors as the rows of the matrix and any lattice point can be written as

$$\lambda = \zeta G,$$

where ζ is an n-dimensional vector of integers [13]. For example, a generator matrix for the m-dimensional integer lattice (often written as $\mathbb{Z}^m$) is the m-dimensional identity matrix. The lattice points of the m-dimensional integer lattice consist of all the possible m-dimensional vectors of integers.

The conventional method of using a lattice for channel coding is to take a finite subset of lattice points and define a one-to-one mapping between the lattice points and binary vectors that represent the information that is to be

sent over a channel. The goal of using lattice codes for channel coding is to maximize the amount of information that can be conveyed over the channel for a given power constraint. As an example, consider the problem of sending bits over an AWGN channel. One method of addressing this problem is to choose a lattice and then map vectors of bits to a finite subset of the lattice. The lattice points will then represent the real values that are sent over the channel and corrupted by noise. The decoder will receive a noisy sequence of points and attempt to recover the bits by decoding the noisy values to the closest lattice points in Euclidean distance. The decoding region for each lattice point is often referred to as its *Voronoi region* and is defined to be the set of points whose Euclidean distance to the given lattice point is closer than that to any other lattice point.

To illustrate the above concepts, consider the hexagonal lattice defined by the generator matrix

$$\mathbf{G} = \begin{bmatrix} 1 & 0 \\ \frac{1}{2} & \frac{\sqrt{3}}{2} \end{bmatrix}.$$

A pictorial representation of the lattice points that are generated by the aforementioned generator matrix is given in Figure 3.4-4, where a finite subset of the hexagonal lattice is shown. If we take four of the lattice points as our finite subset, then we can define a mapping of this subset to binary vectors of length two. In Figure 3.4-4, we chose four lattice points and arbitrarily assigned two-dimensional bit vectors to the lattice points. A transmitter may then parse a bit string into vectors of length 2 and map each vector to a lattice point. Each of the lattice points represent a two-dimensional real vector that will be corrupted by additive white Gaussian noise. The Voronoi region of each lattice point is shown as a hexagon. Therefore, if the additive noise is not large enough to perturb a lattice point outside of its Voronoi region, then the decoder will be able to successfully decode the bits sent by the encoder.

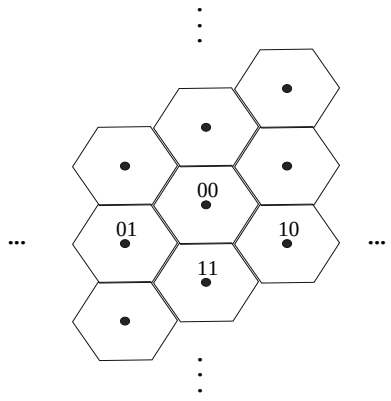


Figure 3.4-4 Example of a hexagonal lattice. A finite subset of the lattice points is shown, and the Voronoi region of each lattice point is a hexagon.

3.4.3.2 Convolutional codes

A class of codes that are often used with both Hamming distance decoders and Euclidean distance decoders are convolutional codes. For simplicity, we restrict our description to binary convolutional codes. Like an (n,k) linear block code, an (n, k) convolutional code maps length-k blocks of information symbols into length-n blocks of output symbols, but each output block depends on the current and previous information blocks. A convolutional code can in general be defined by a linear finite state machine (LFSM). For a binary (n, k) convolutional code of memory ν, the LFSM can be expressed as ν stages of k shift registers that are connected by n different modulo-2 adders, as in Figure 3.4-5. At each time instant, k bits are shifted in to the LFSM and n bits are output from the LFSM. The shift registers in combination with the modulo-2 adders serve to constrain the possible output sequences. The goal of designing a convolutional code is to constrain the possible output sequences to be separated by a large distance. For example, if the convolutional code is to be used with a Hamming distance decoder, then it is desirable to design the convolutional code so that the possible output bit sequences are separated by a large Hamming distance. However, if the convolutional code is to be used with a Euclidean distance decoder, then a mapping between the output bit sequences and vectors of real values must be defined and it is desirable to design the convolutional code so that the possible vectors of real values are separated by a large Euclidean distance. In general, better convolutional codes can be found as the memory is increased.

As is common with LFSMs, it is often beneficial to express the LFSM as a state transition diagram. The states represent the contents of the registers in the LFSM, and the transitions between states are determined by the input bits. As an example, consider the convolutional code shown in Figure 3.4-6a. The parameters of the code are given as $\nu = 2$, $k = 1$, and $n = 2$. In the example, bits are shifted into the registers one at a time and the input bit is represented as the variable u. The contents of the registers that represent the state of the convolutional code are given as variables s_1 and s_2. The output bits of the convolutional code are given as variables c_1 and c_2. We can represent the convolutional code as a state diagram by assigning a circle to each possible state (as in Figure 3.4-6b) and representing the transitions between states with arrows. As can be seen from Figure 3.4-6 the input bit, in combination with the current state of the convolutional code, will determine

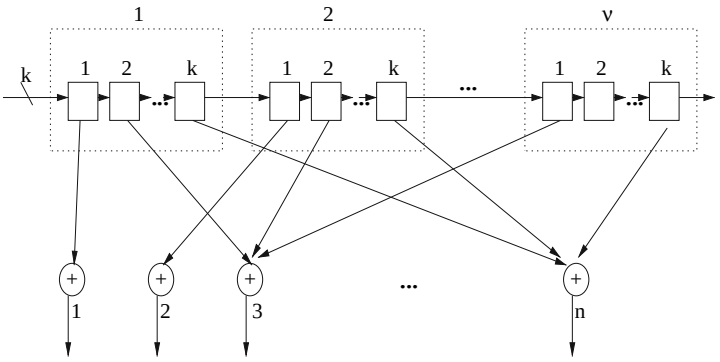


Figure 3.4-5 A linear finite state machine representation of a convolutional code. There are *k* input bits, *n* output bits, and the memory is *v*.

the following state of the convolutional code. For example, if at a given time instant the state of the registers is given as 01 and the input bit is 1, then 1 will be shifted out of the right-most register and the input bit will be shifted into the left-most register. As a result, the following state of the convolutional code will be 10.

Any convolutional code can also be expressed as a vector-matrix product

$$\vec{\mathbf{c}}(D) = \vec{\mathbf{u}}(D)\mathbf{G}(D), \tag{3.4.10}$$

where $\vec{\mathbf{c}}(D) = [c_1(D), c_2(D), \ldots, c_i(D), \ldots, c_n(D)]$ is a row vector of n polynomials, with the ith polynomial representing the ith output bit sequence. A bit sequence, $\{b_0, b_1, \ldots, b_m\}$ can be represented as a polynomial $b_0 + b_1 D + \ldots + b_m D^m$ by weighting the ith bit in time by D^i where D is a variable representing delay. Similarly, $\vec{\mathbf{u}}(D) = [u_1(D), u_2(D), \ldots, u_j(D), \ldots, u_k(D)]$ is a row vector of k polynomials, with the jth polynomial representing the jth bit sequence. The matrix $\mathbf{G}(D)$ is a $k \times n$ matrix that contains generator polynomials that specify the relationship between the input polynomials and the output polynomials. For example, consider the convolutional code shown in Figure 3.4-6a. The first output bit is the modulo-2 addition of the current input bit and the previous input bit in time. We can write a polynomial equation for the first output bit as

$$c_1(D) = u(D) + Du(D),$$

where the variable D represents delay. The second output bit is equal to the modulo-2 addition of the current input bit, the bit from the previous time instant, and the bit from two time instants ago. We can write a polynomial equation for the second output bit as

$$c_2(D) = u(D) + Du(D) + D^2u(D). \tag{3.4.11}$$

Now, the two equations just given can be combined into the form of (3.4.10) where the generator matrix can be expressed as

$$\mathbf{G}(D) = [1 + D \quad 1 + D + D^2]$$

and $\vec{\mathbf{c}}(D) = [c_1(D), c_2(D)]$, $\vec{u}(D) = [u(D)]$. The aforementioned representation of a convolutional code is often useful for analyzing the performance characteristics of a code.

Another representation of a convolutional code is as a trellis. A pictorial representation of a trellis can be formed by aligning all of the possible states in a vertical

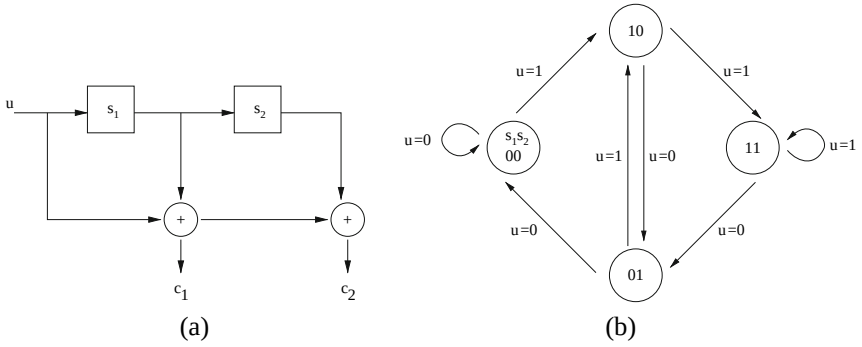


Figure 3.4-6 (a) Convolutional code example with parameters $v = 2$, $k = 1$, and $n = 2$. (b) State diagram representation of convolutional code.

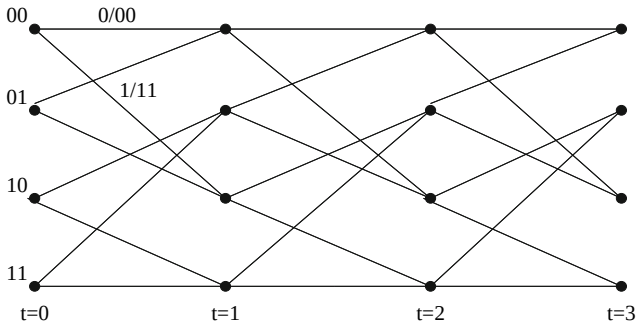


Figure 3.4-7 Trellis representation of convolutional code given in Figure 3.4-6a.

column for each time instant and then connecting the states in accordance with the state transition diagram. The trellis representation of a convolutional code is particularly useful for decoding, as quick decoding algorithms such as the Viterbi decoding algorithm can be derived from the trellis representation. An example of a trellis representation of the convolutional code in Figure 3.4-6a is given in Figure 3.4-7 where the states are represented as dots and the transitions are labeled by the input bit that causes the transition and the resulting output bits.

The goal of the decoder is to find the codeword from the convolutional code closest to the received sequence in either Hamming distance or Euclidean distance (if the output bits of the convolutional code are mapped to real values). This can be done efficiently by using the trellis diagram. If we let $y^{(i)}$ represent the received block (of length n) at time instant i and let $c^{(i)}_{u_i,s_i\to s_{i+1}}$ represent the output block (of length n) at time instant i corresponding to the transition between states s_i and s_{i+1} that results from input block u_i (of length k), then mathematically the goal of the decoder is to find the output sequence $\widehat{\mathbf{c}}$ that is closest in distance to the received sequence,

$$\widehat{\mathbf{c}} = \arg\min_{\mathbf{c}\in\mathcal{C}} \mathbf{d}(\mathbf{c},\mathbf{y}).$$

In the equation just given, $\mathcal{C}$ represents the set of valid codewords, $\mathbf{c} = \{c^{(0)}_{u_0,s_0\to s_1}, c^{(1)}_{u_1,s_1\to s_2}, \ldots, c^{(m)}_{u_m,s_m\to s_{m+1}}\}$ represents a valid sequence of output blocks, and $\mathbf{y} = \{y^{(0)}, y^{(2)}, \ldots, y^{(m)}\}$ represents a sequence of received blocks. Furthermore, $\mathbf{d}(\mathbf{c}, \mathbf{y})$ is the distance metric between the output sequence and the received sequence and can be written as a summation of distances between the received blocks and the output blocks at the various time instants,

$$\mathbf{d}(\mathbf{c},\mathbf{y}) = \sum_{i=0}^{m} d\left(c^{(i)}_{u_i,s_i\to s_{i+1}}, y^{(i)}\right). \tag{3.4.12}$$

Assuming that the trellis starts in state zero (i.e., all of the registers of the convolutional code are cleared to zero), a naive approach to finding the codeword sequence that is closest to the received sequence would be to calculate the distance of the received sequence to each path in the trellis that starts in state zero and then declare the path that is closest to the received sequence as the decoded codeword. This method is inefficient because the amount of computation grows exponentially with the length of the sequence. A more efficient decoding algorithm can be realized by using the Viterbi algorithm. The Viterbi algorithm begins at the first stage of the trellis by calculating the distance between all branches of the trellis that emerge from state zero and the corresponding block in the received sequence (i.e., $d(c^{(0)}_{u_0,s_0\to s_1}, y^{(0)})$). In general, there will be Z^k branches that emerge from any state, so the Viterbi algorithm starts by calculating Z^k distance metrics. For the next stage in the trellis, we can prune paths that end in the same state s_1. More specifically, for all paths that converge to the same state, we can keep the path that has the minimum distance up to that state and prune all other paths. This works because if multiple paths converge to the same state, then any path that may emerge from this state will have an associated distance that will be added to the distance associated with the path that ends in that state. Mathematically, we can break up the total distance metric for any path that goes through a state at time i as the distance metric from state 0 at time 0 to state s_i at time i and the distance metric from state s_i to state s_{m+1} at the end of the trellis:

$$\mathbf{d}(\mathbf{c},\mathbf{y}) = \sum_{j=0}^{i-2} d\left(c^{(j)}_{u_j,s_j\to s_{j+1}}, y^{(j)}\right) + d\left(c_{u_{i-1},s_{i-1}\to s_i}, y^{(i-1)}\right) + \sum_{j=i}^{m} d\left(c^{(j)}_{u_j,s_j\to s_{j+1}}, y^{(j)}\right). \tag{3.4.13}$$

From (3.4.13), we see that all paths that merge at state s_i will have the same possible distances $\sum_{j=i}^{m} d(c^{(j)}_{u_j,s_j\to s_{j+1}}, y^{(j)})$ added to the existing distance of the path and therefore a path with a larger distance at state s_i cannot achieve a smaller overall distance than a path with a smaller distance at state s_i. As a result, we can prune the total number of paths to be no larger than the total number of states. In other words, at each time instant, at most Z^{vk} paths are kept (one path for each state). At time instant m, the minimum-distance path can be determined and traced back to state 0 at time instant 0. The output sequence associated with the minimum distance path is the decoded codeword.

Because the time complexity of the Viterbi algorithm is exponential in the memory order, faster but suboptimal sequential decoding algorithms (e.g., the Fano and Stack algorithms [50]) are used in many time-critical applications. A generalization of the standard Viterbi algorithm is the list Viterbi algorithm (LVA)

[44,41], which finds the L most likely paths instead of only the most likely one.

Symbol MAP decoding (see Section 3.4.2.2) of convolutional codes can be done with soft-input soft-output algorithms. Two of the most prominent ones are the BCJR algorithm of Bahl and colleagues [4] and the soft-output Viterbi algorithm of Hagenauer and Hoeher [25]. Both algorithms output for each information bit u_i, $i = 0,\ldots, km$, an *a posteriori* log-likelihood ratio (LLR)

$$\Lambda(u_i) = \log \frac{Pr\{U_i = 1|\mathbf{Y} = \mathbf{y}\}}{Pr\{U_i = 0|\mathbf{Y} = \mathbf{y}\}}$$

whose sign specifies the reconstructed source bit $\hat{u}_i$.

A family of convolutional codes can be generated from a single convolutional code, called a mother code, with rate $\frac{1}{n}$. Some output symbols of the mother encoder are punctured, which allows the construction of a family of codes with rates $\frac{p}{np}, \frac{p}{np-1}, \ldots, \frac{p}{p+1}$, where p is the puncturing period. To obtain RCPC codes [24], all protection symbols of the higher rate-punctured code are used by the lower rate codes (the higher rate codes are embedded into the lower rate codes). A nice feature of RCPC codes is that if a higher rate code does not provide enough protection, one can switch to a lower rate code simply by adding extra redundant symbols. Another good feature of RCPC codes is that the same Viterbi trellis can be used for all rates.

As mentioned earlier, a convolutional code may be used as either a Hamming distance code or a Euclidean distance code. If the convolutional code is used as a Euclidean distance code, then a mapping between the possible output bits at any given time instant and a set of real values must be defined. One method of defining a mapping is to first choose a constellation of real values such as a finite subset of lattice points and then define a bijective mapping between the lattice points and the possible output bit vectors. For example, the convolutional code shown in Figure 3.4-6a has two output bits, which can assume one of four possible two-bit combinations, so we can define a mapping between the four possible two-bit combinations and the four lattice points shown in Figure 3.4-4. Recall, however, that the goal of code design is to maximize the minimum distance between possible output sequences, and the aforementioned procedure may not maximize the minimum distance for a given convolutional code and a given set of constellation points. A proper Euclidean distance code design should jointly consider the convolutional code and the set of constellation points in defining the mapping between bits and constellation points. This concept was first introduced by Ungerboeck [48], and the resulting codes are often referred to as trellis-coded modulation (TCM) codes.

TCM codes are usually formed by letting a convolutional code index a partition of constellation points [48]. Forney [20] and Conway and Sloane [13] independently utilized this heuristic to define a set of codes that are derived from a convolutional code that indexes a lattice partition. More specifically, both Forney and Conway and Sloane showed that good trellis codes can be obtained by partitioning a well-known lattice and then searching for a convolutional code to index the partition. In this chapter, we will denote a lattice partition as Λ/Λ', where Λ' is a sublattice of Λ and partitions Λ into cosets of Λ'. A coset of Λ' is formed by choosing a lattice point, $\lambda \in \Lambda$, and adding this element to all of the lattice points in Λ'. We denote the coset as $\Lambda' + \lambda$. For example, consider the lattice partition $\mathbb{Z}/4\mathbb{Z}$, where $\mathbb{Z}$ is the integer lattice formed by the 1×1 identity matrix and $4\mathbb{Z}$ is a sublattice of $\mathbb{Z}$ that is formed by scaling the integer lattice by 4. Four disjoint cosets may be formed from $4\mathbb{Z}$ by adding the lattice points $\{0,1, 2, 3\}$ to $4\mathbb{Z}$. Notice that the union of the four cosets is equal to $\mathbb{Z}$. One method of arriving at the four cosets of $4\mathbb{Z}$ is to use a partition tree, $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}$. The first level of the tree is a partition of $\mathbb{Z}$ into two cosets that consist of the even and odd lattice points of $\mathbb{Z}$ (i.e., $2\mathbb{Z}$ and $2\mathbb{Z}+ 1$). The next level of the partition tree further partitions $2\mathbb{Z}$ into $4\mathbb{Z}$ and $4\mathbb{Z}+ 2$ and partitions $2\mathbb{Z} + 1$ into $4\mathbb{Z} + 1$ and $4\mathbb{Z} + 3$. A pictorial representation of the partition tree, $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}$, is given in

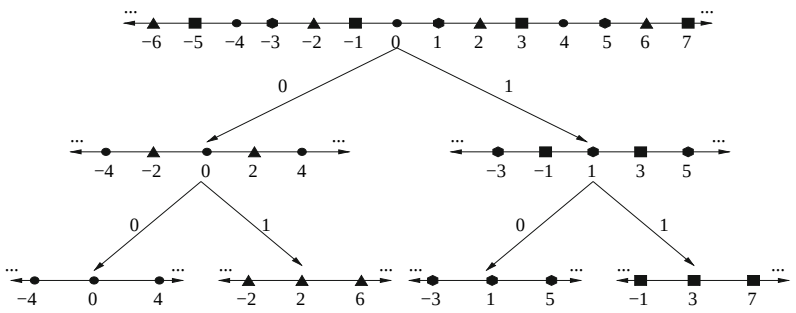


Figure 3.4-8 An example of the partition tree $\mathbb{Z}/2\mathbb{Z}/4\mathbb{Z}$. The branches of the tree are labeled by either 0 or 1 and represent a mapping from bits to cosets of $4\mathbb{Z}$.

Figure 3.4-8. Each of the branches of the partition tree is labeled as either 0 or 1. This labeling defines a mapping between two-bit vectors and cosets of $4\mathbb{Z}$. For example, the coset $4\mathbb{Z}+1$ corresponds to the bit label 10. Now, if we allow the output of a rate-$\frac{1}{2}$ convolutional code to index the labeling of the partition tree for each time instant, then a trellis code may be formed from $\mathbb{Z}/4\mathbb{Z}$ by searching all rate-$\frac{1}{2}$ convolutional codes of a given constraint length to find the convolutional code that maximizes the minimum Euclidean distance between codewords. Note that in the aforementioned, a finite subset of the lattice points must be used to form the trellis code to ensure that there is no ambiguity in decoding a sequence of lattice points to a bit sequence.

The performance of a TCM code is measured by the signal-to-noise ratio (SNR) that is needed to achieve a given probability of error. For high SNRs, it has been shown that the probability of error for a TCM code can be approximated as

$$P_e \approx K_{\min} Q\left(\sqrt{\frac{d_{\min}^2}{4\sigma_N^2}}\right), \tag{3.4.14}$$

where Q is the Q function, $d_{\min}$ represents the minimum distance of the TCM code and $K_{\min}$ represents the number of codewords that have a distance of $d_{\min}$ from a given codeword. We use σ_N^2 to represent the variance of the channel noise. Though effective, the performance of TCM codes is considerably worse than information theoretic bounds.

3.4.3.3 Interleaving

While the codes described earlier are convenient for memoryless channels with small error rates, most of them are not suited to the protection against errors that occur in bursts. When errors occur in bursts, as in fading channels, a transmitted codeword is either free of errors or contains a large number of successive errors. The problem of burst errors can be alleviated with special codes (e.g., Fire codes [17]). An alternative is interleaving, which shuffles the symbols from different codewords before transmission. When a long burst error occurs, the erroneous symbols are distributed among many codewords where they appear as short burst errors. In block interleaving (Table 3.4-2), the channel codewords are placed in the rows of an array, and the codeword symbols are sent columnwise. In cross (or convolutional) interleaving, as shown in Figure 3.4-9, a set of ordered shift registers with linearly increasing memory size is used to separate the output symbols of the channel encoder.

Table 3.4-2 Block interleaver of size 4×7. To transmit four codewords of length 7, the codeword symbols are sent columnwise, in the order 1, 8, 15, 22, …, 7, 14, 21, 28. A burst error of length four produces no more than a single error in a transmitted codeword.

1. Codeword	1	2	3	4	5	6	7
2. Codeword	8	9	10	11	12	13	14
3. Codeword	15	16	17	18	19	20	21
4. Codeword	22	23	24	25	26	27	28

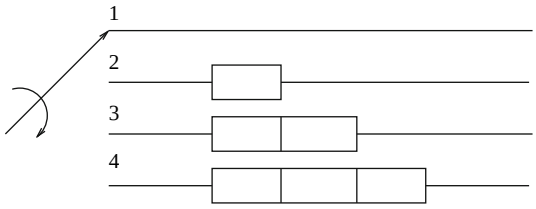


Figure 3.4-9 Cross interleaver with four shift registers. The memory sizes of the shift registers are 0, m, $2m$, and $3m$, respectively. At time unit i, a symbol is inserted into shift register i, which outputs its right-most symbol. Suppose that $m = 1$ and the input symbols are 0, 1, 2, 3, 4, 5, 6, 7, 8,…. After interleaving, the symbols are sent in the order 0, 4, 1, 8, 5,….

3.4.3.4 Turbo codes

In 1993, Berrou *et al.* [9,10] amazed the coding community by introducing a novel class of error-correcting codes, turbo codes, which, for a binary-input AWGN channel, achieved a BER of 10^{-5} with code rate 1/2 and E_b/N_0 as close as 0.5 dB to the practical Shannon limit (see Section 3.4.2.4).

A turbo code is a parallel concatenation of two or more codes connected by pseudo-random interleavers. The constituent codes are usually identical, recursive systematic convolutional (RSC) codes of rate 1/2. An example of an RSC encoder is shown in Figure 3.4-10. Its main property is the existence of a feedback in the shift-register realization.

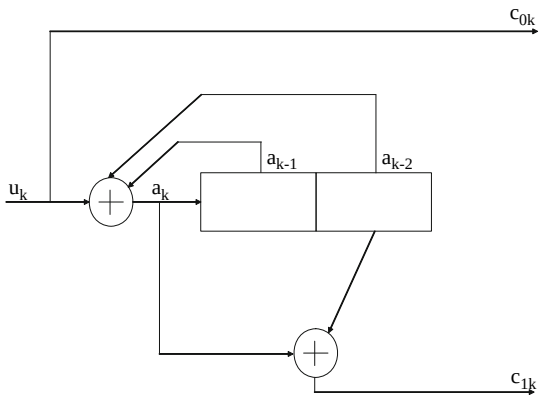


Figure 3.4-10 A recursive systematic convolutional code.

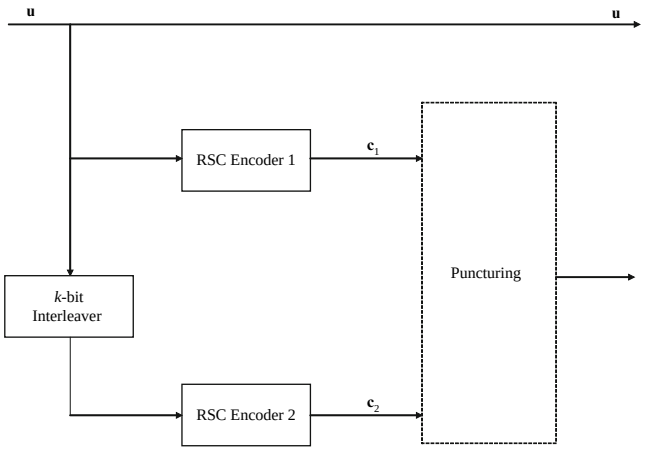


Figure 3.4-11 A classical turbo encoder with two RSC codes.

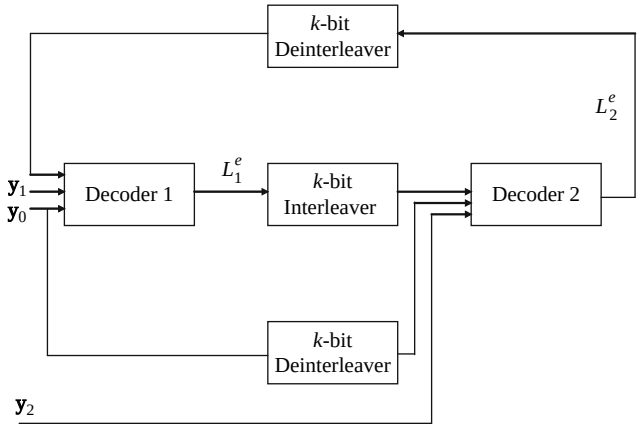


Figure 3.4-12 Block scheme of a classical turbo decoder.

Figure 3.4-11 shows a classical turbo encoder with two constituent RSC codes. In contrast to a serial code concatenation where the output of one encoder forms the input for the next one, in a parallel concatenation, both encoders operate on the same input block. In Figure 3.4-11, an input information block of length k bits, $\mathbf{u} = (u_1, \ldots, u_k)$, is encoded by the first RSC encoder; at the same time, it is passed through a k-bit interleaver and fed into the second RSC encoder. For each input bit u_i, the output consists of that bit u_i and the two parity-check bits $c_{1,i}$ and $c_{2,i}$ from the two RSC encoders. The output corresponding to the input block $\mathbf{u}$ is the codeword $\mathbf{c} = (u_1, \ldots, u_k, c_{1,1}, \ldots, c_{1,k}, c_{2,1}, \ldots, c_{2,k})$. The code rate of the turbo code is 1/3. Higher code rates can be obtained by puncturing the output bits of the two encoders [1]. For example, rate 1/2 can be obtained by alternately puncturing the parity bits of the two RSC encoders. A turbo code is essentially a block code, thus encoding can be seen as a multiplication of the information block by a generator matrix.

One of the many novelties in the turbo code realization is the existence of a block interleaver between the two RSC coders. The interleaver introduces randomness to the code while leaving enough structure in it so that decoding is physically feasible. The size of the interleaver (the length of the information block) is usually very large (in the order of thousands of bits) to ensure good performance. If the size is large enough, any pseudo-random interleaver will perform well. However, for short interleaver sizes, the performance of the code can be significantly enhanced by a clever design of the interleaver [5].

A typical turbo decoder consists of two soft-input soft-output decoders (see Section 3.4.3.2), two k-bit interleavers identical to the encoder interleaver, and a deinterleaver, as shown in Figure 3.4-12. The decoding is based on the symbol **MAP** rule (see Section 3.4.2.2). The a posteriori LLRs for the information bits $u_1, \ldots, u_k$ are estimated in an iterative way by exchanging information between the two constituent decoders. Suppose that the systematic part of the codeword, $\mathbf{c}_0 = (u_1, \ldots, u_k)$, is received as $\mathbf{y}_0$, while the two parity parts, $\mathbf{c}_1 = (c_{1,1}, \ldots, c_{1,k})$ and $\mathbf{c}_2 = (c_{2,1}, \ldots, c_{2,k})$, are received as $\mathbf{y}_1$ and $\mathbf{y}_2$, respectively. In the first iteration, the first decoder generates a reliability information $L_1^e(i)$ for each information bit u_i, $i = 1,\ldots, k$, based on its input, $(\mathbf{y}_0, \mathbf{y}_1)$. This soft-decision output, called extrinsic information, is interleaved and fed to the second decoder. Using its input ($\mathbf{y}_2$ and the interleaved version of $\mathbf{y}_0$), the second decoder computes a reliability information $L_2^e(i)$ for each information bit. Next, the extrinsic information from the first decoder, the extrinsic information from the second decoder, and a channel log-likelihood ratio

$$\log \frac{Pr\{Y_i = y_{0,i} | U_i = 1\}}{Pr\{Y_i = y_{0,i} | U_i = 0\}}$$

are summed to provide a first approximation of the a posteriori LLRs. In the second iteration, the extrinsic information $L_2^e(i)$ is deinterleaved and sent to the first decoder, which exploits this new information to update its extrinsic information. The procedure repeats until the a posteriori LLRs converge or a maximum number of iterations is reached.

Turbo coding with iterative decoding is currently one of the best error-correcting techniques. It significantly outperforms convolutional codes of the same constraint length. One of the key properties of turbo codes is the sharp performance improvement with the increase of the input block length. Thus, to achieve near-capacity performance, large block lengths are needed, which cause huge latency. Therefore, applications of turbo codes are currently limited to those that are not delay sensitive. For example, the new CCSDS telemetry channel coding standard for satellite and deep-space communications uses turbo codes. SMART-1, launched in September

2003 by the European Space Agency, is the first probe that exploits turbo codes. Turbo codes have also been adopted by the leading third-generation (3G) cellular standards, such as CDMA2000 and UMTS.

3.4.4 Hierarchical modulation

Hierarchical modulation [19] is a digital modulation technique that enables transmission of two independent information bit streams with unequal priority on a single channel. As part of the digital terrestrial television standard DVB-T [18], it offers new possibilities in organizing scarce radio frequency bandwidth. In this section, we first outline the main concepts underlying hierarchical modulation and compare it to standard digital nonhierarchical modulation techniques; then, we give examples of possible applications.

Figure 3.4-13 shows constellations of four basic linear digital modulation techniques [38,39]. Each possible digital state (constellation point) in the phase diagram (represented by a dot in Figure 3.4-13) uniquely determines one phase of the carrier signal. Each transmitted bit stream is assigned to one constellation point. The performance of a digital modulation technique can be measured using its achieved data rate (or, equivalently, the number of bits assigned to each digital state) and minimum tolerated SNR for reliable demodulation (which reflects robustness to channel noise). Normally, higher level modulation techniques achieve larger data rates at the expense of a lower robustness.

Binary phase shift keying (BPSK) allows transmission of one bit per modulation signal. The phase of a carrier signal takes two possible values (separated by π) depending on the transmitted bit. 4-Quadrature amplitude modulation (4-QAM), also referred to as Quadrature phase shift keying (QPSK or 4-PSK), transmits two bits on each carrier. Thus, it achieves twice the data rate of BPSK. In Figure 3.4-13, one possible constellation realization is presented, where the carrier phases are $\pi/4$, $3\pi/4$, $5\pi/4$, and $7\pi/4$. In 16-QAM and 64-QAM, because there are 4×4 and 8×8 different constellation points, respectively, four and six bits, respectively, can be sent per modulation signal. The assignment of the bit streams to the digital states is usually determined using Gray-code mapping so that the assignments of the closest constellation points differ in one bit. The data rate is increased compared to 4-QAM at the expense of a lower noise tolerance (due to smaller distances between neighboring states in the phase diagram). For example, compared to 4-QAM with the same code rate, the minimum tolerated SNR is approximately 6 dB and 12 dB higher with 16-QAM and 64-QAM, respectively [19].

Note that in all modulation techniques discussed so far, a single information bit stream (possibly coded) is transmitted per one modulation signal. Hierarchical modulation, however, enables transmission of two

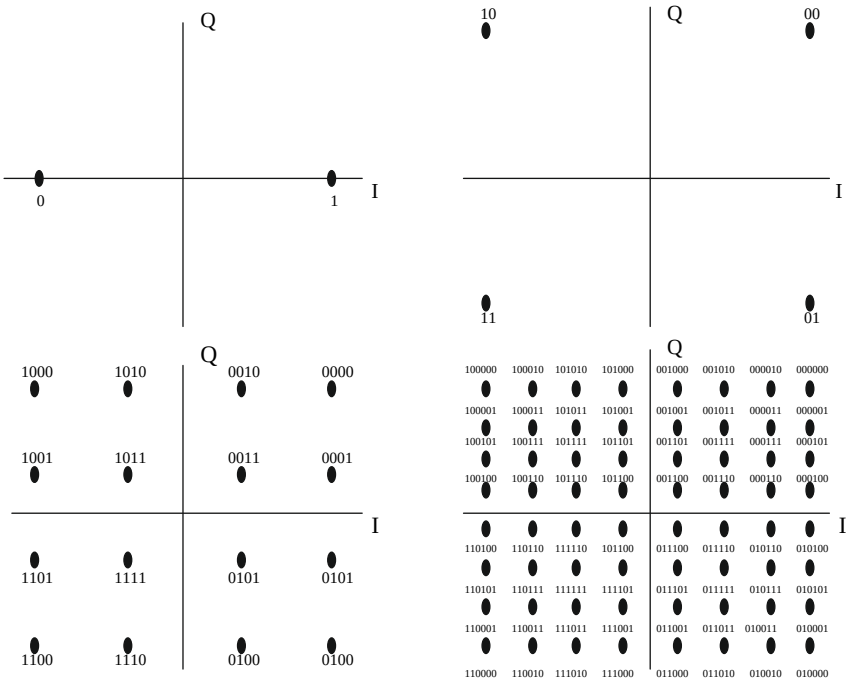


Figure 3.4-13 Constellations of four standard digital modulation techniques: BPSK (top left), 4-QAM (top right), 16-QAM (bottom left), and 64-QAM (bottom right).

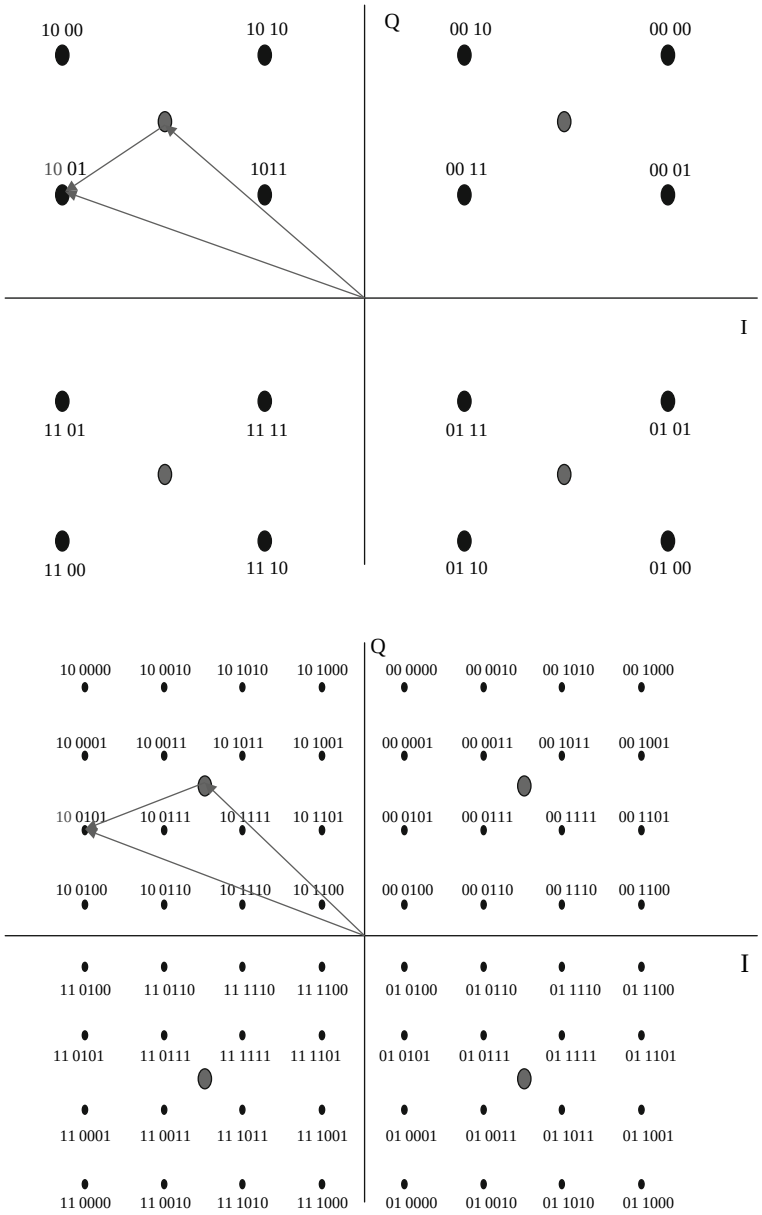


Figure 3.4-14 Hierarchical modulation: "4-QAM in 16-QAM" (top) and "4-QAM in 64-QAM" (bottom).

separate information bit streams in a single modulation signal. One bit stream, called high-priority (HP) bit stream, is embedded within another, called low-priority (LP) bit stream. The main idea is to decouple the bit stream assigned to a digital state into two sub-streams: the first substream is HP, which determines the number of the quadrant (0, 1, 2, or 3) where the digital state is located; the second substream (LP) carries the information about the position of the digital state in the specified quadrant. As a result, hierarchical modulation can be viewed as a combination of 4-QAM (used for the HP bit stream) and either 4-QAM or 16-QAM (used for the LP bit stream).

Two hierarchical modulation constellations are shown in Figure 3.4-14. In the first constellation (the upper figure), 4-QAM is embedded in 16-QAM (thus, it is called "4-QAM in 16-QAM"); in the second one, 4-QAM is embedded in 64-QAM ("4-QAM in 64-QAM"). In both cases, the first two bits constitute the HP bit stream intended for an HP service/client; the remaining two or four bits are the LP bit stream intended for an LP service/client. In the example shown in Figure 3.4-14 (bottom), 10 is sent to the HP clients and 0101 to the LP clients.

Note that the HP bit stream is always modulated as 4-QAM. Thus, as in classic nonhierarchical 4-QAM, it carries two bits per modulation signal. However, because the LP bit stream can be seen at the receiver as an additional noise in the quadrant of the received signal, the HP bit stream is less robust than nonhierarchical

4-QAM (i.e., a higher minimum tolerated SNR is needed).

The LP bit stream is essentially either 4-QAM [Figure 3.4-14 (top)] or 16-QAM [Figure 3.4-14 (bottom)] modulated. Thus, it carries two or four bits and has the same data rate as the corresponding nonhierarchical modulation method. The noise sensitivity is comparable to that of the whole constellation [16-QAM in Figure 3.4-14 (top) or 64-QAM in Figure 3.4-14 (bottom)]. Note that the total rate of the HP and LP bit streams is equal to the rate of the whole nonhierarchical constellation (16-QAM or 64-QAM).

The HP bit stream is obviously more robust to channel noise than the LP bit stream; indeed, a transition of the carrier phase (due to channel noise) from one digital state to the other within a quadrant is more likely to occur than a transition to a state in another quadrant. However, the robustness of the HP and LP bit streams can be further improved by channel coding (i.e., by adding error protection) or by changing the constellation's α factor, as in Figure 3.4-15. The $\alpha = a/b$ factor [18] is defined as the ratio between a, the minimum distance separating two constellation points that carry two different HP bit streams, and b, the minimum distance separating any two constellation points. Constellations with $\alpha > 1$ are called nonuniform constellations. The increase of α makes the HP bit stream more robust at the expense of a less robustness of the LP bit stream. (The DVB-T standard uses $\alpha \in \{1, 2, 4\}$.) Thus, hierarchical modulation splits the actual communication channel in two virtual channels whose characteristics depend on the whole constellation (64-QAM), α factor, and code rates of the HP and LP bit streams.

Hierarchical modulation was originally proposed to enable two different coverage areas for a given transmitter in digital terrestrial TV. It offers great design flexibilities and simplifies network planning. Its value has become even more apparent with recent increasing demands for delivery of different services over heterogeneous networks, where communication channels between the sender and the clients are extremely diverse in available bandwidths and channel noise.

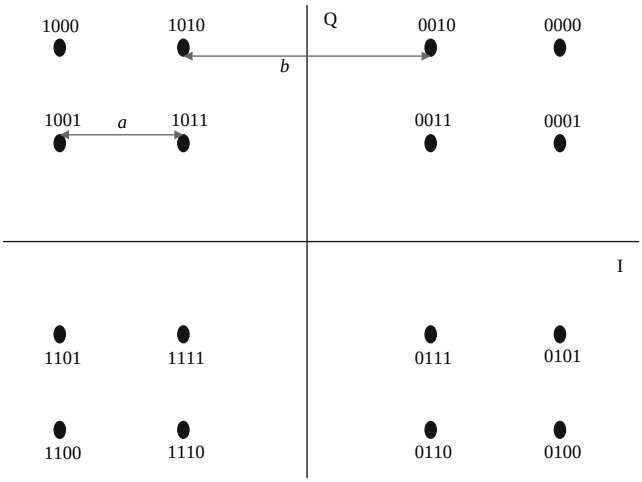


Figure 3.4-15 A nonuniform 16-QAM constellation with $\alpha = b/a = 2$.

For example, suppose that two digital TV programs are to be transmitted simultaneously. With nonhierarchical modulation, the two programs must be broadcast over two separate frequency channels: 4-QAM can be used for the first channel (achieving a data rate of two bits per modulation signal) and 16-QAM for the second channel (with a data rate of four bits per modulation signal). With hierarchical modulation ("4-QAM in 64-QAM"), only one channel is needed: the first program can be transmitted as an HP bit stream (at a data rate of two bits per modulation signal), while the second TV program can be transmitted as an LP bit stream (at a data rate of four bits per modulation signal). Then, the coverage radius (which is determined by the noise tolerance) of the second TV program will be roughly the same as in the nonhierarchical case; the coverage radius of the first program, however, will be smaller than with nonhierarchical 4-QAM, but can be enlarged by increasing the α factor (at the expense of a smaller coverage radius of the second TV program) or by using error protection (at the expense of decreasing the information rate). Thus, one immediate advantage of hierarchical modulation over a nonhierarchical one is the savings in transmission channels because two streams with different data rates and different coverage areas can be transmitted on a single frequency channel.

Hierarchical modulation efficiently addresses the problem of heterogeneity in clients' available bandwidths, receiver resolution capabilities, and channel conditions. For example, a single frequency channel can be used to broadcast a video bit stream to mobile (or portable) receivers and fixed receivers. The mobile receivers will decode the HP bit stream, whereas the fixed receivers will be able, in addition, to decode the LP bit stream (due to their large roof top antenna gains).

Hierarchical modulation can be combined with quality/resolution scalable video coders. Then, the LP bit stream plays the role of the enhancement layer, which improves the quality/resolution of the HP (base layer) bit stream. Depending on transmission conditions, the receiver will be able to decode at the higher or lower quality/resolution level.

Another application is simulcast of the high definition TV formats, together with the standard definition formats. (Transmitting the standard definition together with the high definition formats is necessary because all the receivers do not have screens that support the latter formats.) Here, the HP bit stream carries the standard definition TV formats, and thus will be available to all

receivers, whereas the LP bit stream carries the high definition TV formats only.

Comparisons between hierarchical and non-hierarchical modulation in different scenarios can be found in [42].

3.4.5 Automatic repeat request, hybrid FEC/ARQ

In this section, we present error protection techniques that use retransmissions. Here we assume the presence of a feedback channel from the receiver to the transmitter. We first describe pure ARQ techniques, which are based on error detection and retransmission of the corrupted packets. Then we explain type I hybrid ARQ protocols that combine error correction coding and ARQ techniques. Finally, we overview type II hybrid -ARQ protocols where the transmitter answers a retransmission request by sending additional parity symbols.

3.4.5.1 Pure ARQ protocols

In a pure ARQ system, an information block of length k is encoded into a channel codeword of length n with an error-detecting code. The codeword is sent over the channel and the received word is decoded. If no errors are detected, the transmitted codeword is assumed to be received correctly and need not be retransmitted. Otherwise, the codeword must be sent again until it is received correctly. To send feedback information to the transmitter, the receiver can use a positive acknowledgment (ACK) to indicate that the codeword was received correctly or a negative acknowledgment (NACK) to indicate a transmission error. The efficiency of an ARQ scheme is measured by its reliability and throughput. The reliability is the probability that the receiver accepts a word that contains an undetectable error. The throughput is the ratio of the average number of bits successfully accepted per unit of time to the total number of bits that could be transmitted per unit of time [29]. In the following, we overview the most important ARQ schemes. Details can be found in [29] and [50].

3.4.5.1.1 Stop-and-wait ARQ

In stop-and-wait ARQ, the transmitter sends a codeword and waits for an acknowledgment for that codeword. If an ACK is received, the next codeword is sent. If an NACK is received, the same codeword is retransmitted until it is received correctly, as in Figure 3.4-16. Stop-and-wait ARQ has a simple implementation. In particular, the codewords are not numbered. Its major drawback is the idle time spent by the transmitter waiting for an ACK.

3.4.5.1.2 Go-back-*N* ARQ

In go-back-N ARQ, the transmitter sends the codewords continuously without waiting for an acknowledgment. Suppose that the acknowledgment for codeword c_i arrives after codewords $c_i,\ldots, c_{i+N-1}$ have been sent. If this acknowledgment is of the ACK type, the transmitter sends codeword c_{i+N}. Otherwise, the codewords $c_i,\ldots, c_{i+N-1}$ are sent again, as in Figure 3.4-17. On the receiver side, when an error is detected in a received word, this word and the $N-1$ subsequently received ones are ignored. Note that a buffer for N codewords is required at the transmitter side.

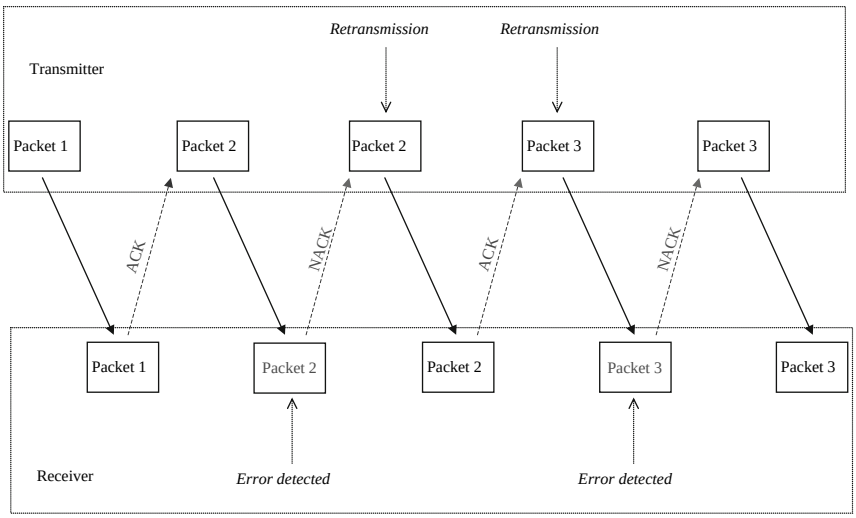


Figure 3.4-16 Stop-and-wait ARQ.

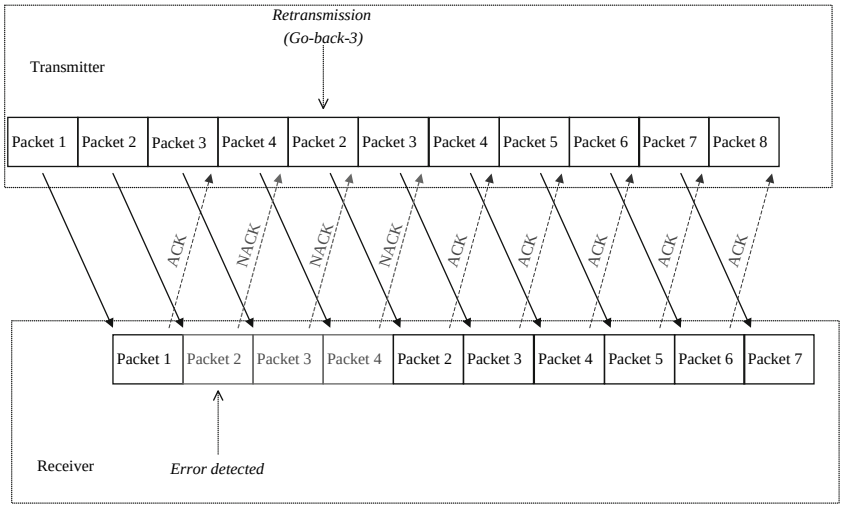


Figure 3.4-17 Go-back-*N* ARQ.

3.4.5.1.3 Selective-repeat ARQ

Selective-repeat ARQ is similar to go-back ARQ. The difference is that when an NACK for codeword c_i is received, only a is retransmitted before the transmission proceeds where it stopped, as in Figure 3.4-18. In addition to the N-codeword buffer at the transmitter, a buffer is needed at the receiver so that the decoded codewords can be delivered in the correct order. This buffer must be large enough to avoid overflow. Selective-repeat ARQ with a finite-size buffer is presented in [29]. An alternative is to combine selective-repeat ARQ with go-back-N ARQ [29]. Here the transmitter switches from selective-repeat ARQ to go-back-N ARQ whenever μ retransmissions of a codeword have been done without receiving an ACK. It switches back to selective-repeat ARQ as soon as an ACK is received. In this way, the buffer size of the receiver can be limited to $\mu(N-1)+1$.

3.4.5.2 Hybrid ARQ protocols

FEC and ARQ can be combined to provide for channels with high error rates better reliability than FEC alone and larger throughput than ARQ alone.

3.4.5.2.1 Type-I hybrid ARQ protocols

In a type-I hybrid ARQ system, each information block is encoded with a channel code with error detecting and error correcting capabilities. This can be a single linear code (see Section 3.4.3.1) or a concatenation of an error detection code as an outer code and an error correction code as an inner code. If the received word can be

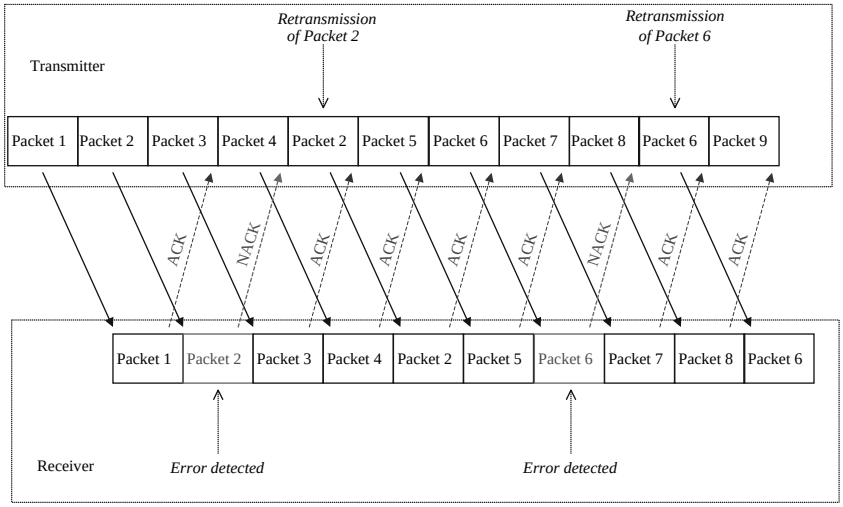


Figure 3.4-18 Selective-repeat ARQ.

correctly decoded, then the decoded codeword is accepted. Otherwise, a retransmission is requested for the codeword.

3.4.5.2.2 Type-II hybrid-ARQ protocols

The basic difference between a type-I hybrid ARQ protocol and a type-II hybrid ARQ protocol is that in the latter the transmitter sends additional parity bits instead of the whole codeword when it receives a retransmission request for this codeword. The following example [50] illustrates the method. An (n, k) MDS code $\mathcal{C}$ is used to encode the information block. The resulting codeword is split in two. The first half can be seen as a codeword $\mathbf{c}_1$ from an $(n/2, k)$ code $\mathcal{C}_1$ and the second one as a codeword $\mathbf{c}_2$ from an $(n/2, k)$ code $\mathcal{C}_2$. Here the two codes $\mathcal{C}_1$ and $\mathcal{C}_2$ are obtained by puncturing the code $\mathcal{C}$. The transmitter starts by sending $\mathbf{c}_1$. If the received word $\mathbf{y}_1$ cannot be correctly decoded, a retransmission is requested. The transmitter then sends the codeword $\mathbf{c}_2$, which is received as $\mathbf{y}_2$. The receiver concatenates $\mathbf{y}_1$ and $\mathbf{y}_2$ and uses the stronger code $\mathcal{C}$ to decode the resulting word.

3.4.6 Summary and further reading

The first part of this chapter presented the fundamental results of information theory, which culminate in Shannon's joint source-channel coding theorem. While this theorem is useful in understanding the theoretical performance bounds for the communication of data over an unreliable channel, it does not explain how a practical communication system should be designed. Practical system design should consider source coding, channel control, and modulation. State-of-the-art channel coding techniques are overviewed in the second part of the chapter. The main message is that channel coding techniques, in particular Turbo codes and LDPC codes, have reached a level of maturity that allows them to achieve performance close to the theoretical bounds announced by Shannon. Another important achievement in the area of channel coding is development of the class of digital fountain codes for protection against packet loss. The third part of the chapter discussed hierarchical modulation, an emerging modulation technique for digital video broadcasting. The last part of the chapter gave a brief survey of error control techniques that rely on data retransmission. These techniques, which require a two-way channel, can be used with error detection only or combined with error correcting codes.

We conclude this chapter with suggestions for further reading. A rigorous treatment of source coding can be found in [6] and [17]. Excellent descriptions of modern channel codes are given in [43] and [33]. The best reference for the latest advances in source and channel coding is the *IEEE Transactions on Information Theory*.

References

[1] O. Acikel and W. Ryan. "Punctured turbo-codes for BPSK/QPSK channels," *IEEE Trans. Commun.*, vol. 47, pp. 1315–1323, September 1999.

[2] V. K. Agarwal and A. Ivanov. "Computing the probability of undetected error for shortened cyclic codes," *IEEE Trans. Commun.*, vol. 40, pp. 494–499, March 1992.

[3] R. B. Ash. *Information Theory*, Dover, New York, 1965.

[4] L. Bahl, J. Cocke, F. Jelinek, and J. Raviv. "Optimal decoding of linear codes for minimizing symbol error rate," *IEEE Trans. Inform. Theory*, pp. 284–287, March 1974.

[5] A. Barbulescu and S. Pietrobon. "Interleaver design for turbo codes," *Electronics Letters*, vol. 30, pp. 2107–2108, December 1994.

[6] T. Berger. *Rate Distortion Theory*, Prentice Hall, 1971.

[7] E. R. Berlekamp. *Algebraic Coding Theory*, McGraw-Hill, New York, 1968.

[8] E. R. Berlekamp, R. J. McEliece, and H. C. A. van Tilborg. "On the intractability of certain coding problems," *IEEE Trans. Inform. Theory*, vol. 24, pp. 384–386, May 1978.

[9] C. Berrou and A. Glavieux. "Near optimum error correcting coding and decoding: turbo-codes," *IEEE Trans. Commun.*, vol. 44, pp. 1261–1271, October 1996.

[10] C. Berrou, A. Glavieux, and P. Thitimajshima. "Near Shannon limit error-correcting coding and decoding: Turbo codes," Proc. *IEEE ICC-1993 International Conference on Communications*, pp. 1064—1070, Geneva, Switzerland, May 1993.

[11] S. A. Butman and R. J. McEliece. "The ultimate limits of binary coding for a wideband Gaussian channel," JPL Deep Space Network Progress Report 42-22, pp. 78–80, 1974.

[12] L. N. Childs. *A Concrete Introduction to Higher Algebra*, Springer, New York, 1995.

[13] J. Conway and N. Sloane. *Sphere Packings and Error-Correcting Codes*, Springer- Verlag, New York, 1988.

[14] S.-Y. Chung, G. D. Forney, T. J. Richardson, and R. Urbanke. "On the design of low-density parity-check codes within 0.0045 dB of the Shannon limit," *IEEE Commun. Letters*, vol. 5, pp. 58–60, February 2001.

[15] T. M. Cover and J. A. Thomas. *Elements of Information Theory*, Wiley, 1991.

[16] D. Divsalar, H. Jin, and R. J. McEliece. "Coding theorems for 'turbo-like' codes," *Proc. 36th Allerton Conf. Communication, Control, and Computing*, pp. 201–210, Allerton, Illinois, September 1998.

[17] R. J. McEliece. *The Theory of Information and Coding*, Cambridge University Press, 2002.

[18] ETSI EN 300 744: Digital video broadcasting (DVB); framing structure, channel coding and modulation for digital terrestrial television, June 2004.

[19] ETSI: Digital video broadcasting (DVB); implementation guidelines for DVB terrestrial services; transmission aspects, Technical Report TR 101190, December 1997.

[20] G. Forney. "Coset codes - part 1: Introduction and geometrical classification," *IEEE Trans. Inform. Theory*, vol. 34, pp. 1123–1151, September 1988.

[21] R. G. Gallager. *Low Density Parity-Check Codes*, MIT Press, Cambridge, 1963.

[22] R. Gray. *Entropy and Information Theory*, Springer-Verlag, New York, 1990.

[23] V. Guruswami and M. Sudan. "Improved decoding of Reed–Solomon and algebraic-geometry codes," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1757–1767, September 1999.

[24] J. Hagenauer. "Rate-compatible punctured convolutional codes (RCPC codes) and their applications," *IEEE Trans. Commun.*, vol. 36, pp. 389–400, April 1988.

[25] J. Hagenauer and P. Hoeher. "A Viterbi algorithm with soft-decision outputs and its applications," *Proc. GLOBECOM*, vol. 3, pp. 1680–1686, Dallas, Texas, November 1989.

[26] J. Jiang and K. Narayanan. "Iterative soft decoding of Reed-Solomon codes," *IEEE Commun. Letters*, vol. 8, pp. 244–246, April 2004.

[27] H. Jin, A. Khandekar, and R. McEliece. "Irregular repeat-accumulate codes," *Proc. 2nd Int. Symposium. Turbo Codes*, pp. 1–8, Brest, France, September 2000.

[28] R. Koetter and A. Vardy. "Algebraic soft-decision decoding of Reed–Solomon codes," *IEEE Trans. Inform. Theory*, vol. 49, pp. 2809–2825, November 2003.

[29] S. Lin and D. Costello, Jr. *Error Control Coding*, 2nd ed., Prentice Hall, 2004.

[30] M. Luby. "LT codes," *Proc. 43rd Annual IEEE Symposium on Foundations of Computer Science*, pp. 271–282, 2002.

[31] M. Luby, M. Mitzenmacher, A. Shokrollahi, D. Spielman, and V. Stemann. "Practical loss-resilient codes," *29th ACM Symposium Theory Computation*, pp. 150–159, 1997.

[32] M. Luby, M. Mitzenmacher, A. Shokrollahi, and D. Spielman. "Efficient erasure correcting codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 569–584, February 2001.

[33] D. J. C. MacKay. *Information Theory, Inference and Learning Algorithms*, Cambridge University Press, 2003.

[34] D. J. C. MacKay and R. M. Neal. "Near Shannon limit performance of low density parity check codes," *Electronics Letters*, vol. 32, pp. 1645–1646, August 1996.

[35] D. J. C. MacKay. "Good error-correcting codes based on very sparse matrices," *IEEE Trans. Inform. Theory*, vol. 45, pp. 399–431, March 1999.

[36] M. MacWilliams and N. Sloane. The *Theory of Error-Correcting Codes*, Karlin, North-Holland, 1992.

[37] J. Pearl. "Fusion, propagation, and structuring in belief networks," *Artificial Intell.* vol. 29, pp. 241–288, 1986.

[38] J. G. Proakis and M. Salehi. *Communication Systems and Engineering*, Prentice-Hall, New Jersey, 2002.

[39] T. S. Rappaport. *Wireless Communications*, Prentice-Hall, New Jersey, 1996.

[40] T. J. Richardson and R. Urbanke. "Efficient encoding of low-density parity-check codes," *IEEE Trans. Inform. Theory*, vol. 47, pp. 638–656, February 2001.

[41] M. Röder and R. Hamzaoui. "Fast tree-trellis list Viterbi algorithm," *IEEE Trans. Commun.*, vol. 54, pp. 453–461, March 2006.

[42] A. Schertz and C. Weck. "Hierarchical modulation," *EBU Technical Review*, April 2003.

[43] C. Schlegel. *Trellis Coding*, John Wiley & Sons, 1997.

[44] N. Seshadri and C.-E. W. Sundberg. "List Viterbi decoding algorithms with applications," *IEEE Trans. Commun.*, vol. 42, pp. 313–323, February–April 1994.

[45] C. E. Shannon. "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623-656, 1948.

[46] C. E. Shannon. "Coding theorems for a discrete source with a fidelity criterion," *I.R.E. Nat. Conv. Rec.*, part 4, pp. 142–163, 1959.

[47] A. Shokrollahi. "Raptor codes," *IEEE Trans. Inform. Theory*, vol. 52, pp. 2551–2567, June 2006.

[48] G. Ungerboeck. "Channel coding with multilevel/phase signals," *IEEE Trans. Inform. Theory*, vol. 28, pp. 55–67, January 1982.

[49] N. Wiberg, H.-A. Loeliger, and R. Kötter. "Codes and iterative decoding on general graphs," *Eur. Trans. Telecommun.*, vol. 6, pp. 513–525, September–October 1995.

[50] S. Wicker. *Error Control Systems for Digital Communication and Storage*, Prentice-Hall, New Jersey, 1995.

Chapter 3.5

3.5

Network adaptive media transport

Mark Kalman and Bernd Girod

3.5.1 Introduction

Internet packet delivery is characterized by variations in throughput, delay, and loss, which can severely affect the quality of real-time media. The challenge is to maximize the quality of audio or video at the receiver, while simultaneously meeting bit-rate limitations and satisfying latency constraints. For the best end-to-end performance, Internet media transmission must adapt to changing network characteristics; it must be *network adaptive*. It should also be *media aware*, so that adaptation to changing network conditions can be performed intelligently.

A typical streaming media system comprises four major components that should be designed and optimized in concert:

1. The *encoder application* compresses video and audio signals and uploads them to the media server.
2. The *media server* stores the compressed media streams and transmits them on demand, often serving hundreds of clients simultaneously.
3. The *transport mechanism* delivers media packets from the server to the client for the best possible user experience, while sharing network resources fairly with other users.
4. The *client application* decompresses and renders the video and audio packets and implements the interactive user controls.

The streaming media client typically employs error detection and concealment to mitigate the effects of lost packets. To adapt to network conditions, the server receives feedback from the client, for example, as positive or negative acknowledgments. More sophisticated client feedback might inform about packet delay and jitter, link speeds, or congestion.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007 Elsevier Inc. All rights reserved

Unless firewalls force them to, streaming media systems do not rely on TCP but implement their own, application-layer transport mechanisms. This allows for protocols that are both network adaptive and media aware. A transport protocol may determine, for example, when to retransmit packets for error control and when to drop packets to avoid network congestion. If the protocol takes into consideration the relative importance of packets and their mutual dependencies, audio or video quality can be greatly improved.

The media server can implement intelligent transport by sending the right packets at the right time, but the computational resources available for each media stream are often limited because a large number of streams must be served simultaneously. Much of the burden of an efficient and robust system is therefore on the encoder application, which, however, cannot adapt to the varying channel conditions and must rely on the media server for this task. Rate scalable representations are therefore desirable to facilitate adaptation to varying network throughput without requiring computation at the media server. Switching among bit streams encoded at different rates is an easy way to achieve this task, and this method is widely used in commercial systems. Embedded scalable representations are more elegant and are preferable, if the rate–distortion penalty often associated with scalable coding can be kept small.

Section 3.5.2 begins with a review of the framework for rate–distortion optimized (RaDiO) media streaming initially proposed by Chou and Miao [6]. In the sections that follow, we discuss extensions to the framework. Section 3.5.3 shows how rich acknowledgments can be incorporated. In Section 3.5.4, we discuss the importance of multiple deadlines for video packets. Section 3.5.5 discusses how the framework can be extended to include a more accurate statistical model for characterizing packet

loss and delay. In Section 3.5.6, finally, we discuss an alternative to rate–distortion optimized streaming that directly minimizes congestion instead of rate.

3.5.2 Rate–distortion optimized streaming

We start by reviewing the seminal work by Chou and Miao on rate–distortion optimized (RaDiO) streaming [6]. They considered streaming as a stochastic control problem, with the goal of determining which packets to send and when to minimize reconstruction distortion at the client for a given average transmission rate. Our discussion serves as the starting point for the extensions and variations described in the later sections.

3.5.2.1 Basic RaDiO framework

Let us assume that a media server has stored a compressed audio or video stream that has been packetized into data units. Each data unit l has a size in bytes B_l and a deadline by which it must arrive at the client in order to be useful for decoding. The importance of each data unit is captured by its *distortion reduction* ΔD_l, a value representing the decrease in distortion that results if the data unit is decoded. Often, distortion is expressed as mean-squared error, but other distortion measures might be used as well.

Whether a data unit can be decoded often depends on which other data units are available. In the RaDiO framework, these interdependencies are expressed in a directed acyclic graph. An example dependency graph is shown for SNR-scalable video encoding with Intra (I), Predicted (P), and Bidirectionally predicted (B) frames (Figure 3.5-1). Each square represents a data unit and the arrows indicate the order in which data units can be decoded.

The RaDiO framework can be used to choose an optimal set of data units to transmit at successive transmission opportunities. These transmission opportunities are assumed to occur at regular time intervals. Because of decoding dependencies among data units, the importance of transmitting a packet at a given transmission opportunity often depends on which packets will be transmitted in the near future. The scheduler therefore makes transmission decisions based on an entire optimized plan that includes anticipated later transmissions. Of course, to keep the system practical, only a finite time horizon can be considered.

The plan governing packet transmissions that will occur within a time horizon is called a *transmission policy*, π. Assuming a time horizon of N transmission opportunities, π is a set of length-N binary vectors π_l, with one such vector for each data unit l under consideration for transmission. In this representation, the N binary elements of π_l indicate whether, under the policy, the data unit l will be transmitted at each of the next N transmission opportunities. The policy is understood to be contingent upon future acknowledgments that might arrive from the client to indicate that the packet has been received. No further transmissions of an acknowledged data unit l are attempted, even if π_l specifies a transmission for a future time slot.

Each transmission policy leads to its own *error probability*, $\varepsilon(\pi_l)$, defined as the probability that data unit l arrives at the client late, or not at all. Each policy is also associated with an expected number of times that the packet is transmitted under the policy, $\rho(\pi_l)$. The goal of the packet scheduler is to find a transmission policy π with the best trade-off between expected transmission rate and expected reconstruction distortion. At any transmission opportunity the optimal π minimizes the Lagrangian cost function

$$J(\pi) = D(\pi) + \lambda R(\pi), \tag{3.5.1}$$

where the expected transmission rate

$$R(\pi) = \sum_l \rho(\pi_l) B_l, \tag{3.5.2}$$

and the expected reconstruction distortion

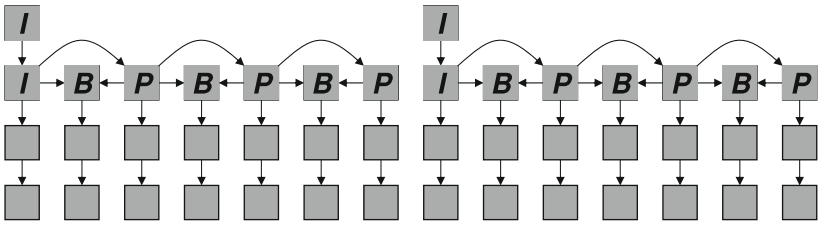


Figure 3.5-1 A directed acyclic graph captures the decoding dependencies for an SNR-scalable encoding of video with L-frames, P-frames, and B-frames. Squares represent data units and arrows indicate decoding order.

$$D(\pi) = D_0 - \sum_l \Delta D_l \prod_{l' \preceq l} (1 - \varepsilon(\pi_{l'})). \quad (3.5.3)$$

The Lagrange multiplier λ controls the trade-off between rate and distortion. In (3.5.3) D_0 is the distortion if no data units arrive, ΔD_l is the distortion reduction if data unit l arrives on time and can be decoded, and the product term $\prod_{l' \preceq l}(1 - \varepsilon(\pi'_l))$ is the probability for this to occur. The notation $l' \preceq l$ is shorthand for the set of data units that must be present to decode data unit l.

In the aforementioned formulation, delays and losses experienced by packets transmitted over the network are assumed to be statistically independent. Packet loss is typically modeled as Bernoulli with some probability, and the delay of arriving packets is often assumed to be a shifted-Γ distribution. Expressions for $\varepsilon(\pi_l)$ and $\rho(\pi_l)$ can be derived in terms of the Bernoulli loss probabilities, the cumulative distribution functions for the Γ-distributed delays, the transmission policies and transmission histories, and the data units' arrival deadlines. These derivations are straightforward, but because the resulting expressions are cumbersome, they are omitted here.

The scheduler reoptimizes the entire policy π at each transmission opportunity to take into account new information since the previous transmission opportunity and then executes the optimal π for the current time. An exhaustive search to find the optimal π is generally not tractable; the search space grows exponentially with the number of considered data units, M, and the length of the policy vector, N [14]. Even though rates and distortion reductions are assumed to be additive, the graph of packet dependencies leads to interactions, and an exhaustive search would have to consider all 2^{MN} possible policies. Chou and Miao's RaDiO framework [6] overcomes this problem by using conjugate direction search. Their *Iterative Sensitivity Adjustment (ISA)* algorithm minimizes (3.5.1) with respect to the policy π_l of one data unit while the transmission policies of other data units are held fixed. Data units' policies are optimized in round-robin order until the Lagrangian cost converges to a (local) minimum.

Rewritten in terms of the transmission policy of one data unit, (3.5.1), (3.5.2) and (3.5.3) become

$$J_l(\pi_l) = \varepsilon(\pi_l) + \lambda' \rho(\pi_l), \quad (3.5.4)$$

where $\lambda' = \frac{\lambda B_l}{S_l}$ incorporates the rate–distortion trade-off multiplier λ from (3.5.1), the data unit size B_l, and S_l, a term that expresses the sensitivity of the overall expected distortion to the error probability $\varepsilon(\pi_l)$ of data unit l. The sensitivity term represents the relative importance of a particular data unit and incorporates the error probabilities of the other data units that l depends on. The sensitivity S_l changes with each iteration of the ISA algorithm to take into account the optimized policy for the other data units.

Figure 3.5-2 demonstrates improved video streaming performance achieved with RaDiO. Luminance PSNR versus transmitted bit rate is plotted for streaming simulations using an H.263+two-layer SNR scalable encoding of the *Foreman* sequence. The frame rate is 10 fps; a Group of Pictures (GOP) consists of one I-frame followed by nine P-frames. The encoded source rate is 32 kbps for the base layer alone and 69 kbps when the enhancement layer is added. The results are for a simulated channel in which packet losses occur independently with a loss rate of 20%, and packet delays are drawn as independent, shifted-Γ random variables with a mean delay of 50 ms and a standard deviation of 25 ms. Figure 3.5-2 plots PSNR versus transmitted bit rate for a heuristic, prioritized ARQ system and for R–D optimized system. In the ARQ system, the client requests retransmissions for packets that do not arrive by a time interval after they are expected, and the server transmits these requested packets with priority as long as the requested packet may still reach the client in time for playout. When the capacity of the channel falls below the source rate for the enhanced stream, the ARQ system sends only the base layer packets. Both the ARQ and the R–D optimized system use an initial preroll delay of 400 ms. By continuously optimizing its packet transmission choices, the optimized system makes use of the SNR and temporal scalability of the source encoding to finely tune the source rate to the available channel capacity, yielding substantial gains.

Several techniques have been proposed to further reduce the complexity of the basic RaDiO algorithm.

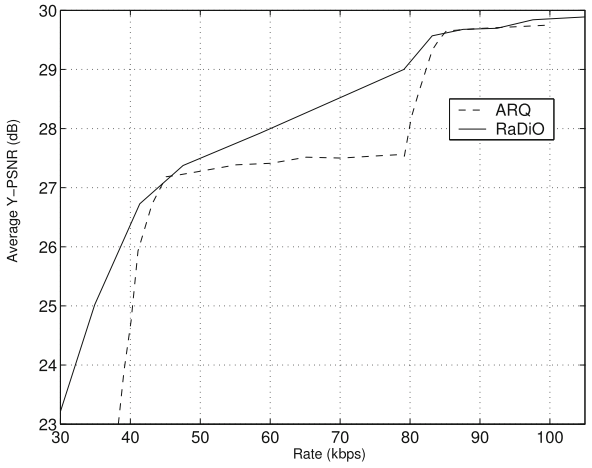


Figure 3.5-2 PSNR vs. transmitted bit rate for a video streaming system that uses heuristic deadline-constrained prioritized ARQ and for a system that uses RaDiO transmission scheduling. The results are for an H.263+ SNR scalable encoding [9] of the *Foreman* sequence.

Chou and Sehgal [7] have presented simplified methods to compute approximately optimized policies. The framework appears to be robust against simplifications of the algorithm and approximations to ΔD_l, the information characterizing the value of individual packets with respect to reconstruction distortion. An attractive alternative to ISA is a randomized algorithm recently developed by Setton in which heuristically and randomly generated candidate policies are compared at each transmission opportunity [15,17]. The best policy from the previous transmission opportunity is one of the candidates and thus past computations are efficiently reused. With a performance similar to ISA, the randomized algorithm usually requires much less computation.

3.5.2.2 Receiver-driven streaming

When transmitting many audio and video streams simultaneously, a media server might become computation limited rather than bandwidth limited. It is therefore desirable to shift the computation required for network adaptive media transport from the server to the client to the extent possible. Fortunately, rate–distortion optimized streaming can be performed with the algorithm running at the client so that very little computation is required at the server [7].

For *receiver-driven* streaming, the client is provided information about the sizes, distortion reduction values, and interdependencies of the data units available at the server ahead of time. The size of this *hint track* or *rate–distortion preamble* is small relative to the media stream. The receiver uses this information to compute a sequence of requests that specify the data units that the server should transmit. It is straightforward to adapt the algorithm discussed in Section 3.4.2.1 to compute a sequence of requests that yield an optimal trade-off between the expected transmission rate of the media packets that the server will send and the expected reconstruction distortion that will result [7]. Figure 3.4-3 illustrates the differences between sender-driven and receiver-driven streaming.

By combining sender-driven and receiver-driven techniques, the RaDiO framework can be applied to diverse network topologies. For example, RaDiO might be implemented in a proxy server placed between the backbone network and a last hop link (Figure 3.5-4) [3]. The proxy coordinates the communication between the media server and the client using a hybrid of receiver- and sender-driven streaming. End-to-end performance is improved compared to a sender- or receiver-driven RaDiO system because traffic created by retransmissions of media packets lost in the last hop to the client does not need to traverse and congest the backbone link.

3.5.3 Rich acknowledgments

In one extension to the RaDiO framework, streaming performance is improved through the use of *rich acknowledgments* [4]. In sender-driven RaDiO streaming using conventional acknowledgments, when a client receives a media packet, the client sends an acknowledgment packet (ACK) to the server. If the ACK packet is lost, the server may decide to unnecessarily retransmit the packet at the expense of other packets.

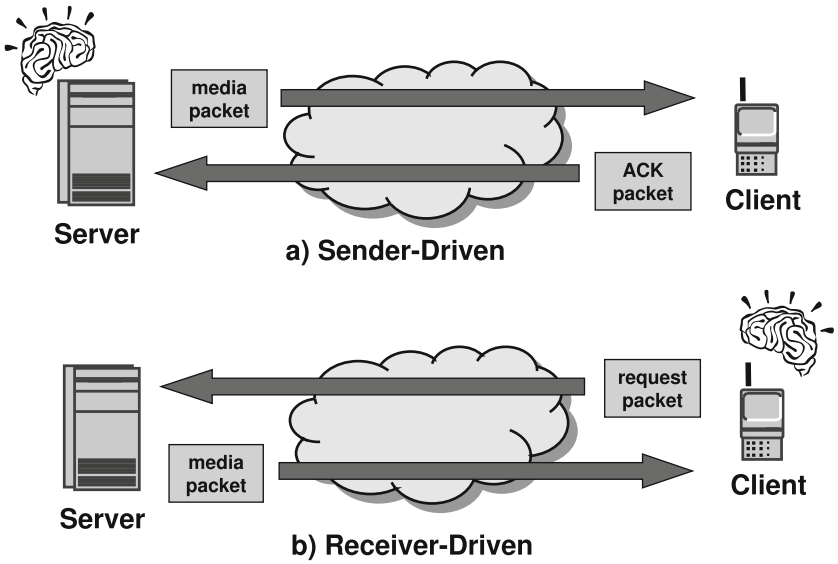


Figure 3.5-3 In sender-driven streaming (a), the server computes an optimal sequence of media packet transmissions and the client acknowledges packets upon receipt. In receiver-driven streaming (b), the complexity is shifted to the client. The client computes an R–D optimized sequence of requests to send to the server and the server only needs to respond to the client's requests.

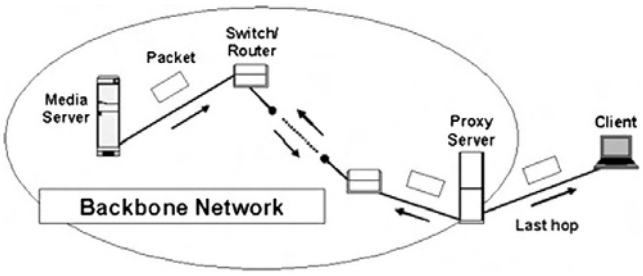


Figure 3.5-4 Proxy-driven RaDiO streaming. A proxy server located between the backbone network and a last hop link uses a hybrid of receiver- and sender-driven RaDiO streaming to jointly optimize requests to send to the server and media packets to forward to the client.

With rich acknowledgments, the client does not acknowledge each data unit separately. Instead, it periodically transmits a packet that positively acknowledges all packets that have arrived so far and negatively acknowledges (NACK) packets that have not yet arrived. A rich acknowledgment packet hence provides a snapshot of the state of the receiver buffer.

Rich acknowledgments require some changes to the basic RaDiO framework described in Section 3.5.2. As shown in [6], a transmission policy π_l for a data unit can be understood in terms of a Markov decision process. At discrete times t_i the server makes an observation o_i and then takes a transmission action a_i specifying *send* or *don't send*. Sequences of observation and action pairs (o_i, a_i) in time can be enumerated in a Markov decision tree. Each node q_i in the tree specifies a particular history of observations and actions (a_0, o_0), (a_1, o_1), …, (a_i, o_i). A transmission policy specifies what transmission action will be taken as a function of what state q_i is reached in the tree.

A Markov decision tree is shown in Figure 3.5-5. The tree enumerates the possible sequences of observation–action pairs for the transmission of a data unit using the rich acknowledgment scheme. In the tree, possible actions a are *send* or *don't send*. Possible observations o are (Ø), no relevant feedback has arrived, *ACK*, a feedback packet has acknowledged the reception of the data unit, or *NACK*, a feedback packet has indicated that the packet has not been received by the feedback packet's send time. NACKs with different time stamps are distinct observations. In contrast, in the conventional feedback scheme in which each packet is acknowledged individually upon receipt, there are only two possible observations, *ACK* and Ø. Regardless of the scheme, the optimization algorithm calculates the probabilities of each path through the tree given a policy and then chooses the policy that yields the best trade-off between expected number of transmissions $\rho(\pi_l)$ and loss probability $\varepsilon(\pi_l)$.

Figure 3.5-6 compares average PSNR versus transmitted bit rate for the 13-s *Foreman* sequence streamed using the rich feedback scheme and using the conventional acknowledgment scheme. Two-layer SNR-scalable H.263+ is used for the encoding [9]. The bit rate of the base layer alone is 32 kbps with an average PSNR of 27 dB. When the enhancement layer is added, the encoded rate becomes 69 kbps with an average PSNR of 30.5 dB. The results are for simulation experiments with a 10% loss rate for both media packets and feedback packets.

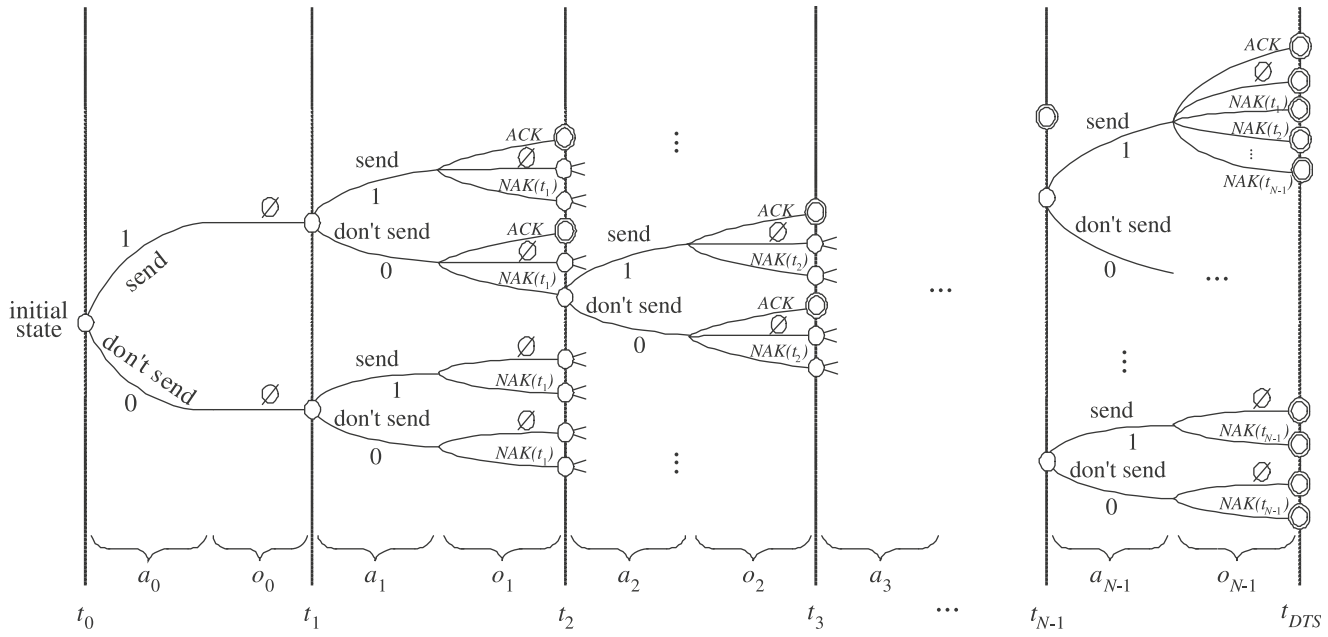


Figure 3.5-5 State space for the Markov decision process when *rich acknowledgments* are used.

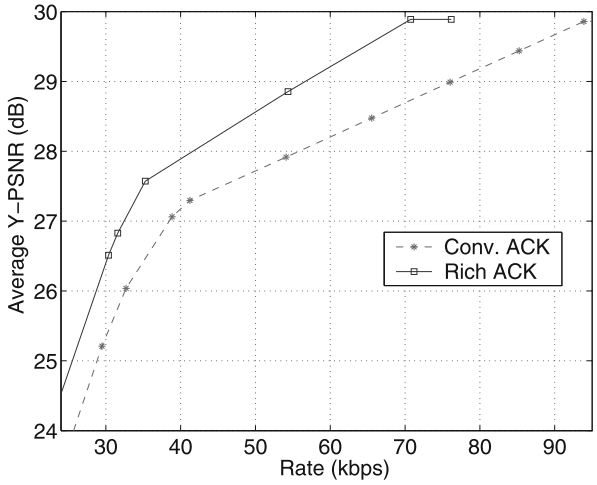


Figure 3.5-6 Rich vs. conventional acknowledgments for rate–distortion optimized streaming of QCIF Foreman.

Delays for packets not lost were distributed according to independent shifted-Γ distributions with shift $\kappa = 50$ ms, mean $\mu = 25$ ms, and standard deviation $\sigma = 35$ ms.

In Figure 3.5-6, the rich acknowledgment scheme outperforms the RaDiO scheme with conventional ACKs for all transmission rates. The maximum PSNR improvement is 1.3 dB at a transmitted bit rate of 70 kbps. The improved performance of the rich acknowledgment scheme is due to the robust transmission of the feedback information. With rich acknowledgments, the effect of a lost feedback packet is mitigated because subsequent feedback packets contain the same (updated) information. In addition, because rich acknowledgment packets also provide NACKs, there is less ambiguity for the server to resolve. In the case of conventional feedback, a nonacknowledged transmission may be due to a lost media packet or to a lost acknowledgment packet.

3.5.4 Multiple deadlines

In the RaDiO framework described in Section 3.5.2, ΔD_l is the expected distortion reduction if data unit l is decodable by its deadline. It was assumed that a data unit l must arrive by its specific deadline in order for its distortion reduction ΔD_l to be realized and in order for data units dependent on l to be decoded. Late data units are discarded. Often, however, a data unit arriving after its deadline is still useful for decoding.

As an example, consider the case of bidirectional prediction with a sequence of frames I-B-B-B-P. In the RaDiO framework, the deadline for the P-frame would be determined by the decoding time of the first B-frame. If the P-frame arrives later, however, it should not be discarded. It may still be useful for decoding subsequent B-frames or at least for decoding and displaying the P-frame itself. Thus there are several deadlines associated with the P-frame, each with its own associated distortion reduction [18].

Another example where a data unit may be associated with multiple deadlines is the case of decoders that allow *Accelerated Retroactive Decoding* (ARD). This idea was initially proposed in the context of MPEG-2 transmission over ATM [8]. ARD makes use of the ability of many streaming clients to decode video faster than real time. With ARD, when late-arriving data units finally do arrive, the decoder goes back to the frames corresponding to the late-arriving packets and quickly again decodes the dependency chain up to the current playout time, but now without error. In this way the remaining pictures in the GOP can be decoded and displayed without degradation.

As shown in [12], the introduction of multiple deadlines results in changes to expressions used to calculate expected distortion $D(\pi)$ for R–D optimized streaming. For each data unit, there is no longer a single error probability, but a set of them, one for each of the frame deadlines associated with that data unit. This results in changes to (3.5.4) that express the Lagrangian cost (3.5.1) as a function of only the transmission policy of one data unit π_l while other policies are held fixed. With multiple deadlines, the expression in (3.5.4) becomes

$$J_l(\pi_l) = \rho(\pi_l) + \sum_{i \in W_l} v_{t_i} \varepsilon(\pi_l, t_i), \tag{3.5.5}$$

where W_l is the set of frames that require data unit l for decoding, i is the frame index, and t_i is the decoding deadline for frame i. The quantity $\varepsilon(\pi_l, t_i)$ is the probability that data unit l does not arrive by deadline t_i. As before, $\rho(\pi_l)$ is the expected number of times data unit l is transmitted under policy π_l.

In (3.5.5) the quantity v_{t_i}, given $v_{t_i} = \frac{S_{l,t_i}}{\lambda B_l}$, by is analogous to the reciprocal of λ' in (3.5.4). Note that the sensitivity term S_{l,t_i} is also indexed by the deadline. It is the sensitivity of the overall distortion to the arrival of data unit l by deadline t_i.

Figure 3.5-7 shows performance gains due to the multiple deadline formulation in the case when ARD is implemented in a streaming video client. PSNR-versus-rate results are shown for the *Foreman* sequence streamed in a low-latency application in which the preroll delay is 100 ms. In the simulation experiments, frames of video were due for decoding 100 ms after they became available for transmission at the server. The packet loss rate was 20% in both directions, and delays for packets not lost were independent, identically distributed (iid) shifted-Γ with shift $\kappa = 10$ ms, mean $\mu = 40$ ms, and standard deviation $\sigma = 23$ ms. The sequence was encoded using a two-layer SNR-scalable H.263+ [9], at 10 fps with prediction structure I-P-P-P… and GOP

length of 20 frames. The base and enhancement layer bit rates and PSNRs were similar to those of the sequence in Section 3.5.3.

In Figure 3.5-7, PSNR-versus-rate curves compare the multiple-deadline schemes and the single-deadline scheme, as well as a heuristic scheme. The heuristic scheme uses prioritized, deadline-limited ARQ in which base layer retransmissions had highest priority, followed by base layer transmissions, enhancement layer transmission, and enhancement layer retransmissions. Retransmissions were triggered when packets were not acknowledged within the 0.90 point of the cumulative distribution function of the round-trip time. Figure 3.5-7 shows that the multiple-deadline formulation yields up to a 3-dB improvement over a single deadline. The single-deadline scheme does not recognize the utility of late packets and often misses opportunities to schedule valuable data units close to, or after, their original deadlines. The R–D optimizing schemes outperform the heuristic schemes regardless of whether the heuristic schemes are used with ARD-enabled clients.

3.5.5 Dependent packet delays

In the R–D optimized streaming algorithms discussed in Sections 3.5.2, 3.5.3, and 3.5.4, the delays of successive packets have been modeled as iid shifted-Γ random variables with loss also occurring independently as described in [6]. The iid model simplifies calculations for $\varepsilon(\pi_l)$, the error probability due to a transmission policy, and for $\rho(\pi_l)$, the expected number of transmissions that will result from a transmission policy. It fails to capture the dependence among delays, however. In the Internet, successive packets usually travel along the same path, might experience a similar backlog while waiting in the same queues, and rarely arrive out of order. This results in strongly dependent delays of successive packets.

In streaming simulations that employ measured Internet delay traces, we have observed that the iid model can lead to suboptimal scheduling performance. For example, Figure 3.5-8 shows simulation results when packets were delayed according to a delay trace measured over a 14-hop Internet path with a cable modem last hop, as described in [10]. At transmission rates above 80 kbps, the multidead-line R–D optimizing formulation described in Section 3.5.4 is outperformed by the simple heuristic ARQ scheme (also described in Section 3.5.4). The suboptimal performance at high rates is due to the iid delay model assumed by the R–D optimization algorithm. With the iid model, policies that specify repeated transmission of a data unit at successive opportunities yield lower calculated error probabilities for errors due to late loss. The algorithm mistakenly believes that if the data unit is delayed the first time it is transmitted, subsequent transmissions may arrive earlier and on time. Thus at higher rates, the algorithm sends packets multiple times even though in our measured trace the loss probability is very low (0.014%) and packets always arrive in the order they are transmitted.

Rate–distortion performance can be improved by modeling packet delays at successive transmission time slots as a first-order Markov random process [10]. In [11] we have presented an R–D optimization scheme that uses this model. In the scheme, feedback packets inform the server about the delay over the channel in the recent

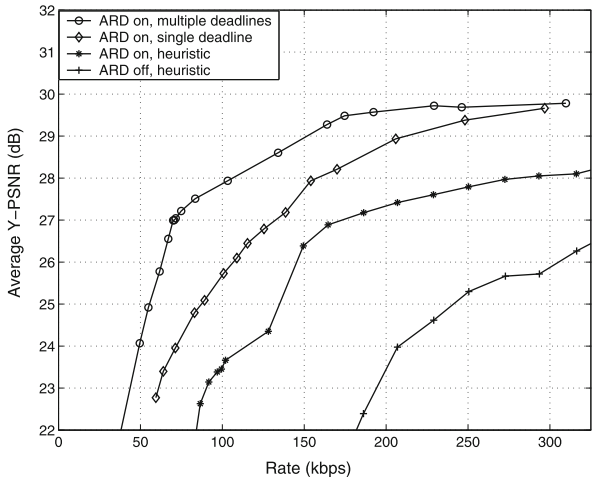


Figure 3.5-7 Rate–distortion performance of schedulers for the Foreman sequence streamed over a simulated channel with iid shifted-Γ-distributed packet delays and 20% Bernoulli loss. End-to-end latency $d = 100$ ms. A PSNR improvement of up to 3.15 dB is observed for the optimizing scheduler that considers multiple deadlines compared to the one that considers a single deadline.

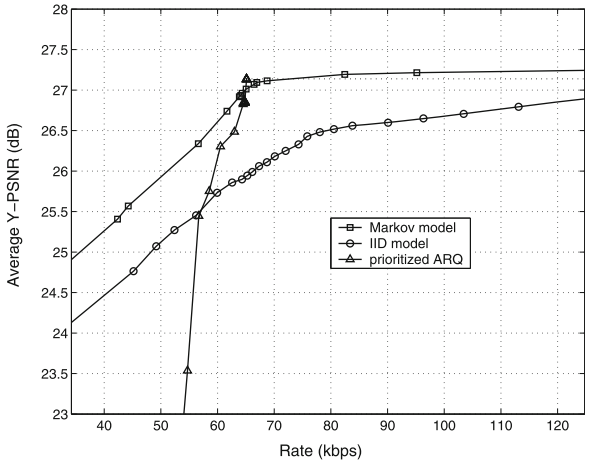


Figure 3.5-8 Rate–distortion performance of schedulers for the *Foreman* sequence streamed over a measured Internet delay trace. End-to-end latency $d = 150$ ms. The RaDiO scheduler that models delays as iid is suboptimal at high rates where it is outperformed by a heuristic-prioritized ARQ scheduler. The scheduler that models delays as a first-order Markov random process yields PSNR improvement of up to 1.1 dB over the iid scheduler.

past. Using this feedback and a family of conditional delay distributions, the scheme can more accurately calculate the expected distortion $D(\pi)$ and the expected transmission rate $R(\pi)$ resulting from a transmission policy π.

Figure 3.5-8 shows that the RaDiO scheme using the Markov channel model outperforms the RaDiO scheduler that uses iid delay modeling by up to 1.1 dB and is not outperformed by the heuristic scheduler at low rates. We note that the mean PSNR for all experiments is limited because the delays in the 14-hop cable modem trace are often greater than the 150-ms preroll. Because the client uses the ARD scheme discussed in Section 3.5.4 and because the packet loss rate is nearly zero, the heuristic scheme, which uses time-out triggered retransmissions with the time-out set to 2 · (estimated RTT), performs nearly optimally at high transmission bit rates.

3.5.6 Congestion–distortion optimized scheduling

RaDiO streaming and its various extensions described do not consider the effect that transmitted media packets may have on the delay of subsequently transmitted packets. Delay is modeled as a random variable with a parameterized distribution; parameters are adapted slowly according to feedback information. In the case when the media stream is transmitted at a rate that is negligible compared to the minimum link speed on the path from server to client, this may be an acceptable model. In the case where there is a bottleneck link on the path from server to client, however, packet delays can be strongly affected by *self-congestion* resulting from previous transmissions.

In [16] a congestion–distortion optimized (CoDiO) algorithm is proposed, which takes into account the effect of transmitted packets on delay. The scheme is intended to achieve an R–D performance similar to RaDiO streaming but specifically schedules packet transmissions in a way that yields an optimal trade-off between reconstruction distortion and congestion, measured as average delay, on the bottleneck link. As with RaDiO, transmission actions are chosen at discrete transmission opportunities by finding an optimal policy over a time horizon. However, in CoDiO, the optimal policy minimizes the Lagrangian cost $D + \lambda\Delta$, where D is the expected distortion due to the policy Δ and A is the expected end-to-end delay, which measures congestion.

CoDiO's channel model assumes a succession of high-bandwidth links shared by many users, followed by a bottleneck last hop used only by the media stream under consideration. CoDiO needs to know the capacity of the bottleneck, which can be estimated, for example, by transmitting back-to-back packets [13]. The overall delay experienced by packets is captured by a gamma pdf that is dynamically shifted by an extra delay that models the self-inflicted backlog at the bottleneck. Since the bottleneck is not shared and its capacity is known, the backlog can be accurately estimated. This channel model is used to calculate the expected distortion D due to packet loss and the expected end-to-end delay Δ.

The performance of the CoDiO scheme is illustrated using ns-2 simulation experiments [1,16]. The first hop is a high-bandwidth 45-Mbps link with 22-Mbps exponential cross-traffic. The second hop is a 50-kbps link that carries only the video traffic to be scheduled. The video encoding used is the same as that described in Section 3.5.4. The preroll delay for the experiments is 600 ms. Figure 3.5-9 plots luminance PSNR versus average end-to-end delay for the CoDiO and the RaDiO schemes. The various points on the curves were generated by varying λ, which trades-off congestion–distortion in the case of CoDiO and rate distortion in the case of RaDiO. The graphs show that the CoDiO scheme resulted in end-to-end delays that were approximately half of those measured for the RaDiO scheme at the same PSNR. Transmission rates versus PSNR for both schemes are almost identical (Figure 3.5-10).

CoDiO outperforms RaDiO because it distributes transmissions in time and attempts to send packets as late as safely possible. This reduces the backlog in the bottleneck queue and hence the average end-to-end delay. Other applications sharing the network experience less congestion. RaDiO, however, is less network-friendly. As the scheduler only considers average rate, its traffic tends to be more burst, relying more on the buffering in the network itself.

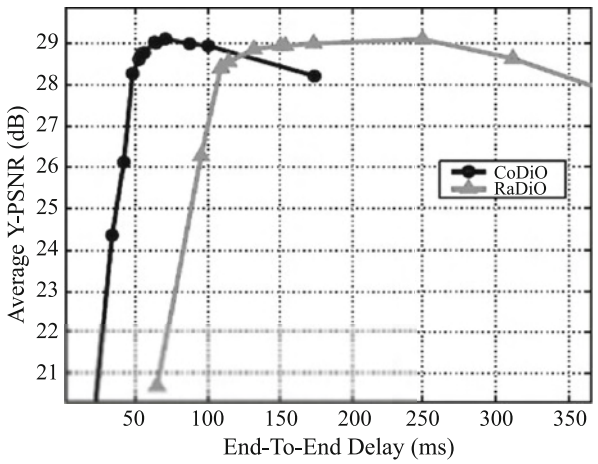


Figure 3.5-9 Performance comparison of RaDiO and CoDiO for video streaming over a bottleneck link. The horizontal axis shows the expected end-to-end delay due to the congestion caused by the video traffic. CoDiO causes much less congestion than RaDiO at the same PSNR. From [16].

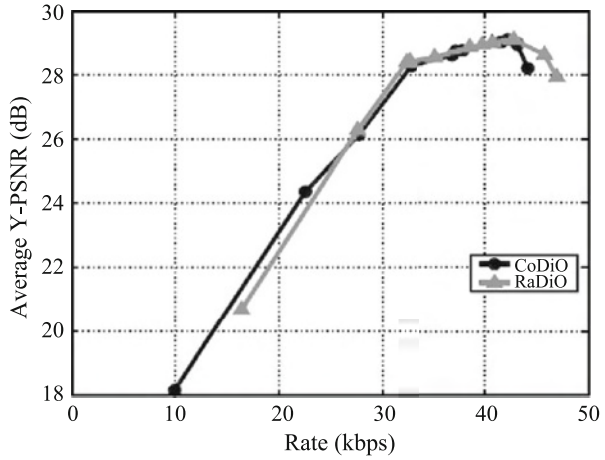


Figure 3.5-10 Rate–distortion performance of RaDiO and CoDiO for video streaming over a bottleneck link. From [16].

3.5.7 Summary and further reading

In this chapter we have discussed network adaptive media transport through the RaDiO framework for rate distortion optimized media streaming. After reviewing the basic framework as initially presented by Chou and Miao, we considered extensions and enhancements that have been proposed. The framework can be implemented in a media server or, alternatively, at the client. Rich acknowledgments are an easy way to improve resilience against losses in the feedback channel. For video streaming, it is useful to incorporate multiple deadlines for packets. Considerable gains are possible by accelerated retroactive decoding of packets, particularly if a multiple-deadline scheduler knows about this client capability and schedules accordingly. RaDiO typically assumes independent packet delays, but, in fact, Internet packet delays are highly dependent. In Section 3.5.5, an extension of RaDiO streaming that utilizes a Markov model of successive packet delays has been shown to rectify the poor performance that arises due to its simple iid channel model. Finally, we have considered self-congestion that might arise with streaming over a bottleneck link. Congestion–distortion optimized streaming, CoDiO, yields the same PSNR performance as RaDiO, but reduces the congestion, measured in terms of end-to-end delay.

Readers with further interest should first study Chou and Miao's seminal paper [6] in depth. The paper is based on a longer technical report [5], so readers might want to consult this document as well. Interestingly, numerous papers appeared during the review period of [6], inspired by this work, many of which are now referenced in [6] itself. Various extensions and the most comprehensive experiments applying RaDiO to streaming of H.264/AVC encoded video can be found in Chakareski's dissertation [2]. The best reference for CoDiO and low-complexity randomized scheduling algorithms so far is Setton's dissertation [15]. There are numerous research groups active in the area, and their publications might be of interest to those following the evolving state-of-the-art in network adaptive media transport.

References

[1] The Network Simulator – ns-2. *www.isi.edu/nsnam/ns/*.

[2] J. Chakareski. *Rate-Distortion Optimized Packet Scheduling for Video Streaming*. Ph.D. thesis, Rice University, Houston, TX, 2005.

[3] J. Chakareski, P. A. Chou, and B. Girod. Rate-distortion optimized streaming from the edge of the network. In *IEEE Workshop on Multimedia Signal Processing*, St. Thomas, U.S. Virgin Islands, December 2002.

[4] J. Chakareski and B. Girod. Rate-distortion optimized video streaming with rich acknowledgments. In *Proceedings SPIE Visual Communications and Image Processing VCIP-2004*, Santa Clara, CA, January 2004.

[5] P. A. Chou and Z. Miao. Rate-distortion optimized streaming of packetized media. Technical report MSR-TR-2001-35, Microsoft Research, Redmond, WA, 2001.

[6] P. A. Chou and Z. Miao. Rate-distortion optimized streaming of packetized media. *IEEE Transactions on Multimedia*, 8(2):390–404, April 2006.

[7] P. A. Chou and A. Sehgal. Rate-distortion optimized receiver-driven streaming over best-effort networks. In *Packet Video Workshop*, Pittsburgh, PA, April 2002.

[8] M. Ghanbari. Postprocessing of late cells for packet video. *IEEE Transactions on Circuits and Systems for Video Technology*, 6(6):669–678, December 1996.

[9] ITU-T. Video coding for low bitrate communication: Recommendation H. 263, Version 2, 1998.

[10] M. Kalman and B. Girod. Modeling the delays of successively-transmitted internet packets. In *IEEE Conference on Multimedia and Expo*, Taipei, Taiwan, June 2004.

[11] M. Kalman and B. Girod. Rate-distortion optimized video streaming using conditional packet delay distributions. In *IEEE Workshop on Multimedia Signal Processing*, Siena, Italy, September 2004.

[12] M. Kalman, P. Ramanathan, and B. Girod. Rate distortion optimized streaming with multiple deadlines. In *IEEE International Conference on Image Processing*, Barcelona, Spain, September 2003.

[13] V. Paxson. *Measurement and Analysis of End-to-End Internet Dynamics*. Ph.D. dissertation, UC Berkeley, Berkeley, CA, 1997.

[14] M. Podolsky, S. McCanne, and M. Vetterli. Soft ARQ for layered streaming media. Technical Report UCB/CSD-98-1024, University of

California, Computer Science Department, Berkeley, CA, November 1998.

[15] E. Setton. *Congestion-Aware Video Streaming over Peer-to-Peer Networks*. Ph.D. dissertation, Stanford University, Electrical Engineering, 2006.

[16] E. Setton and B. Girod. Congestion-distortion optimized scheduling of video over a bottleneck link. In *IEEE Workshop on Multimedia Signal Processing*, Siena, Italy, September 2004.

[17] E. Setton, J. Noh, and B. Girod. Congestion-distortion optimized peer-to-peer video streaming. In *Proc. IEEE International Conference on Image Processing, ICIP-2006*, Atlanta, GA, October 2006.

[18] S. Wee, W. Tan, J. Apostolopoulos, and M. Etoh. Optimized video streaming for networks with varying delay. In *Proceedings of the IEEE International Conference on Multimedia and Expo 2002*, Lausanne, Switzerland, August 2002.

Chapter 3.6

3.6

Real-time communication over networks

Yi Liang, Yen-Chi Lee, and Andy Teng

3.6.1 Introduction

Internet protocol (IP)-based real-time communication, including voice-over IP (VoIP), video telephony, and teleconferencing, has been gaining popularity in recent years. One example is VoIP, which has been competing with the traditional public switched telephone network (PSTN) for years and now enjoys increased market share. This is due to the many advantages of IP-based communication, including lower cost as well as the capability of providing integrated data, voice, and video, a larger variety of features, and more value-added services.

Despite rapid expansion and improvement of the underlying infrastructure, quality-of-service (QoS) is still one of the major challenges of real-time communication over IP networks. The unreliable and stateless nature of today's IP results in a best-effort service, that is, packets may be delivered with an arbitrary delay or may even be lost. Transmitted over the best-effort network and suffering from variable throughput, delay, and loss, data packets have to be delivered by a deadline to become useful. Excessive delay severely impairs communication interactivity; packet loss results in glitches in audio and poor picture quality and frozen frames in video. The heterogeneity of today's Internet also poses a major challenge for media delivery to users with various connection speeds, where scalability is highly desirable.

The challenges that the industry faces, in conjunction with the commercial promise of the technology, have attracted considerable effort in research and product development. In this chapter, we will first describe an architecture for real-time communication, followed by topics on how to improve the QoS. In Section 3.6.2, we will describe the basic system architecture as well as two categories of the most important protocols: signaling and transport. In Section 3.6.3, we will address QoS issues, especially minimizing latency, combating loss, adapting to available bandwidth, and audio–video synchronization.

3.6.2 Architecture and fundamentals

3.6.2.1 Systems

Figure 3.6-1 shows the setup of a typical VoIP system. An IP phone or a sufficiently equipped PC connects to the Internet to be able to make VoIP calls. For traditional phones in the PSTN network, a gateway is needed for the interoperation between the PSTN and the Internet. After introducing signaling and transport protocols later in this section, we will illustrate the process of setting up a call using corresponding protocols in more detail (Figure 3.6-8).

Figure 3.6-2 shows a typical architecture for real-time audio and video communication over IP networks. Transport protocols, including UDP, TCP, and Real-Time Transport (RTP)/Real-Time Transport Control Protocols (RTCP) are built on top of the IP layer. Audio and video codecs are applied on the content encapsulated or to be encapsulated in the payload. The upper-layer applications call audio and video codecs to perform data compression. Signaling protocols, such as session initiation protocol (SIP), are used for call setup and control. The signaling and transport protocols are described in more detail in the following sections.

3.6.2.2 Signaling protocols

The SIP, originally developed by the IETF Multi-Party Multimedia Session Control Working Group, is the most

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007 Elsevier Inc. All rights reserved

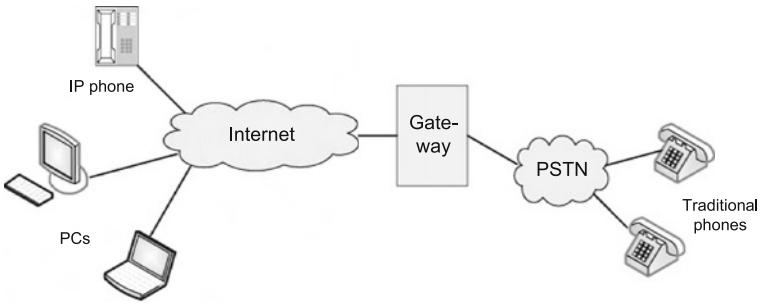


Figure 3.6-1 A typical configuration for a VoIP system with both IP-based devices and traditional PSTN phones.

widely used signaling protocol for real-time conversational applications over IP networks. As a signaling protocol, SIP provides the following functions:

- Call setup and tear down;
- Advanced features, such as call hold, call waiting, call forwarding, and call transfer;
- Capability exchange;
- Interoperability between different types of networks (e.g., PSTN) and different signaling protocols (e.g., H.323);
- Multicasting.

Moreover, SIP has been designed to be scalable enough to support simultaneous calls for a substantial number of users and to be extensible enough to include more features and functions in the future.

SIP may be transported by either TCP or UDP. TCP provides a reliable, connection-oriented transport, while UDP provides a best-effort, connectionless transport across the Internet. Port numbers 5060 and 5061 are the default ports for SIP, although any number above 49172 may be used. The protocol stack for SIP-based IP phone service is shown in Figure 3.6-3.

There are two types of SIP messages: request and response. The request message is initiated by a user agent client (UAC) for registering, call setup, tear down, acknowledgment, etc., while the response message is generated by a user agent server (UAS) or an SIP server in response to the request.

The request message in SIP, as with the other IETF protocols (e.g., RTSP), is called a "method." There are six fundamental SIP methods considered as basic signaling for call setup and tear down, which are defined in IETF RFC 3261 [1]: INVITE, ACK, BYE, REGISTER, CANCEL, and OPTIONS. Specifically, the INVITE method is used to initiate a call. The ACK method is used by the call originator to acknowledge the final response to the INVITE request. The BYE method is used to terminate a call. The REGISTER method is used by a user agent (UA) to register itself to an SIP server with the addressing information (contact URI). The CANCEL method is used to cancel the request sent earlier. The OPTIONS method is used to query an SIP server/client capability. In addition to the six methods defined in RFC 3261, other methods were added later as SIP extensions and specified in different RFCs. Examples include INFO

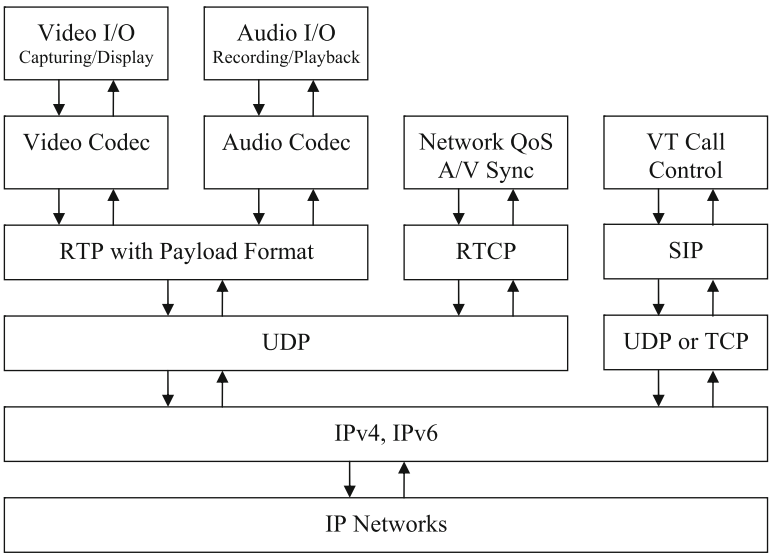


Figure 3.6-2 A typical architecture for an IP-based video and audio communication system.

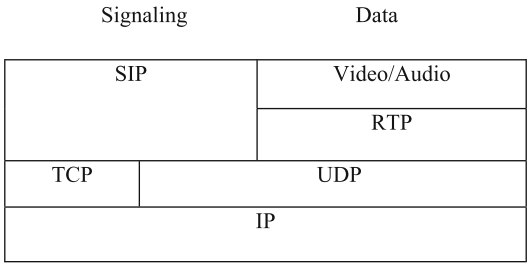


Figure 3.6-3 Protocol stack (signaling flow and data flow).

(RFC 2976 [2]), MESSAGE (RFC 3428 [3]), NOTIFY (RFC 3265 [4]), PRACK (RFC 3262 [5]), REFER (RFC 3515 [6]), SUBSCRIBE (RFC 3265 [4]), and UPDATE (RFC 3311 [7]).

The response message is called the "response code" in SIP. The SIP response codes are inherited from HTTP/1.1, except for the 6xx class, which is defined by SIP itself (RFC 3261). The six classes of SIP response codes are described briefly here.

1xx (provisional response): information to indicate current status before a definitive response. The 1xx response is designed such that an ACK is never triggered by it and thus the reliability for 1xx transmission is not critical. `180 Ringing` is an example of a 1xx response, which is used to inform the originator that the UA has already received the INVITE request.

2xx (successful response): a response used to indicate that the request has been successfully received. Example: `2 00 OK`.

3xx (redirectional response): information used to indicate the user's new location or alternative services.

4xx (request failure): a response used by a UAS or a server to indicate that the request cannot be processed due to authorization failure, authentication failure, account issue, requesting itself, or other problems not related to the server. Example: `400 Bad Request` indicates that the server does not understand the request.

5xx (server failure): a response used by a UAS or a server to indicate that the request cannot be processed due to the server's problem. Examples include `500 Server Internal Error` and `501 Not Implemented`.

6xx (global failure): a response used to indicate that the response will fail in all locations and thus the request should not be delivered.

A simple message flow for call setup and tear down is illustrated in Figure 3.6-4. The SIP request message is composed as follows: method name (e.g., INVITE), address, header fields, and message body. Each response message consists of a code (e.g., 200 OK), header fields, and message body. Note that the header fields and the message body may not appear in all messages.

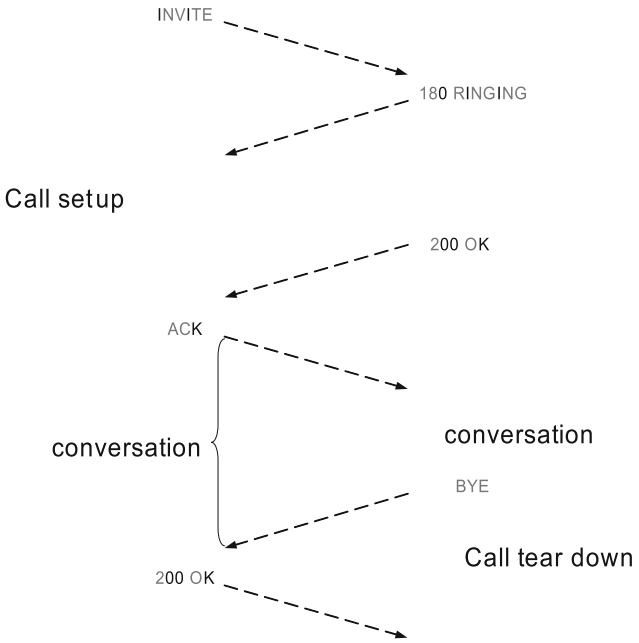


Figure 3.6-4 Simple message flow for call setup and tear down.

3.6.2.2.1 Address

SIP supports a variety of addressing schemes, including SIP URI (Uniform Resource Identifiers), secure SIP URI, telephone URI, and e-mail URL (Uniform Resource Locator). SIP URI is usually represented as `sip:<userinfo>@<host>:<port>`.

3.6.2.2.2 Header fields

A header field is composed as `<header>:<field>`. There are 44 header fields defined in RFC 3261: Accept, Accept-Encoding, Accept-Language, Alert-Info, Allow, Authentication-Info, Authorization, Call-ID, Call-Info, Contact, Content-Disposition, Content-Encoding, Content-Language, Content-Length, Content-Type, CSeq, Date, Error-Info, Expires, From, In-Reply-To, Max-Forwards, Min-Expires, MIME-Version, Organization, Priority, Proxy-Authenticate, Proxy-Authorization Proxy-Require, Record-Route, Reply-To, Require, Retry-After, Route, Server, Subject, Supported, Timestamp, To, Unsupported, User-Agent, Via, Warning, and WWW-Authenticate. The most common ones are introduced here:

Call-ID: used to uniquely identify a call. Example: `CALL- ID : t315fde3-68te-33uyr@test.com`

Contact: used to carry a URI that identifies the resource requested or the request originator. Example: `Contact: sip: johnsmith@test.com`

CSeq: a decimal number used to uniquely identify a request. All responses corresponding to a request use

the same CSeq as the request. The CSeq number is usually increased by one for a new request.

From: used to specify the originator. Example: `From: ''John Smith'' <sip: johnsmith@test.com>`

Max-Forwards: an integer in the range of 0–255 used to specify the maximum number of hops that a message can take. The recommended initial value is 70. It is decreased by one as the message passes through a proxy or gateway. The proxy/gateway discards the message when the value is dropped to zero.

To: used to specify the recipient of the request.

Via: used to record the path the request has been traveled. The response walks through the same path in the reverse order.

The mandatory headers for the six fundamental requests are shown in Table 3.6-1.

3.6.2.2.3 Message body

Although any format can be used as a message body, the Session Description Protocol (SDP) [8] is the most popular one. SDP specifies media information such as media type, codec, author, title, encryption key, bandwidth, start time, and end time. SDP can be used for capability exchange at the call set-up stage. An example of SDP message is shown here

```
m = audio 49170 RTP/AVP 102
a = rtpmap:102 AMR/8000
a = fmtp:102 maxptime=60; octect-
align=1; mode-set=4
m = video 49350 RTP/AVP 110
a = rtpmap:110 MP4V-ES/90000
a = fmtp:110 profile-level-id=0;
config=000001B.....
```

The SDP message just given specifies the following information. Audio is transported by the RTP/AVP protocol through port 49170, with payload number 102. The RTP timestamp resolution is 1/8000 s. The audio is coded by AMR with the maximum bit rate of 7.4 kbps. Three-frame bundling is used, three audio frames are bundled together to form an RTP packet. Audio packetization should follow the rules defined in RFC 3267 [9]. Video is transported by the RTP/AVP protocol through port 49350, with payload number 110. The RTP timestamp resolution is 1/90,000 s. The video is coded by MPEG-4 video SVP L0 (simple visual profile level 0). Video packetization should follow the rules defined in RFC 3016 [10].

SIP is an IP telephony signaling protocol developed by the IETF, which competes with the H.323 protocol developed by the ITU-T for the same application. The fundamental difference between the two protocols is that SIP is a text-based protocol and inherits the rich set of the IETF protocols, such as SDP, whereas H.323 is binary encoded and utilizes many features from other ITU-T protocols, for example, H.245. Comparisons between the two protocols on features, packet loss recovery, security mechanism, and capability exchange are listed in Table 3.6-2. A more detailed comparison can be found in [11].

3.6.2.3 Media transport and control protocols

The commonly used media transport and control protocols in IP voice and video telephony applications are RTP and RTCP, as defined in RFC 3550. RTP and RTCP are designed to be independent of the underlying transport and network layers. Applications usually run RTP and RTCP on top of UDP and IP, as shown in Figure 3.6-3.

3.6.2.3.1 Real-time transport protocols

RTP provides end-to-end delivery services for media data that have real-time characteristics. It defines useful information such as timestamp, sequence number, and marker, to allow receivers to keep the order of the packets, and to play out media at the proper pace. This is due to the

Table 3.6-1 Mandatory header fields for six fundamental SIP methods, where M denotes *mandatory*.

Header/Requests	INVITE	ACK	BYE	REGISTER	CANCEL	OPTIONS
Call-ID	M	M	M	M	M	M
Contact	M					
CSeq	M	M	M	M	M	M
From	M	M	M	M	M	M
Max-Forwards	M	M	M	M	M	M
To	M	M	M	M	M	M
Via	M	M	M	M	M	M

Table 3.6-2 Comparisions between SIP and H.323.

Comparisons	SIP	H.323
Encoding method	Text	Binary
Family	IETF	ITU-T
Transport	TCP or UDP	TCP
Packet loss recovery	Through SIP itself	Through TCP
Capability exchange	SDP (simple but limited)	H.245 (rich but complicated)
Security	Through other IETF; protocols for encryption, authentication, etc.	Not very good
Features	Call holding, call transfer, call forwarding, call waiting, conferencing, instant messaging	Call holding, call transfer, call forwarding, call waiting, conferencing

fact that IP networks often introduce jitter in packets' arrival time and sometimes packet reordering. RTP itself, however, does not provide any mechanism to ensure timely delivery or to provide another quality of service.

Figure 3.6-5 shows the format of an RTP packet and its RTP header. Typically, in one-to-one telephony applications, the size of the RTP header is 12 bytes (no CSRC). V is a 2-bit field that identifies the version of the RTP. P is 1-bit information used to indicate if there are any padding octets at the end that are not part of the payload. X is a 1-bit field used to tell if there is any header extension information. CC means CSRC count, which uses 4 bits in the header and contains the number of CSRC identifiers that follow the header with fixed size. If an RTP session is one to one, such as in a video telephony application, the CSRC count should be set to zero. PT indicates payload type in 7 bits. It tells the format of the payload that an RTP packet carries.

M is a 1-bit marker and its interpretation is defined by a profile or payload format. For example, RFC 3016 is the payload format used for MPEG-4 audio and video. It specifies that if an encoded video frame is carried in multiple RTP packets, the marker bit of the last packet should be set to one to indicate the end of the frame. This is particularly useful for the RTP receiver to signal the video decoder to decode a video frame as soon as the last packet arrives.

SN specifies the sequence number of the RTP packet. It increases by one when one RTP data packet is sent. The initial sequence number of the first RTP packet for an RTP session should be randomly generated. For different media, the initial value may be different. For real-time telephony applications, the receiver can use the sequence numbers to detect any lost packets.

The timestamp TS reflects the sampling time of the first octet in the RTP packet payload. The sampling time should be calculated from a clock that increases monotonically and linearly in time to allow synchronization and jitter calculations. The timestamp may increase at a different pace for different media. For example, speech data are usually sampled at 8000 Hz and each speech frame can typically have 160 samples. Each RTP packet for speech will have a timestamp increment of 160. For video data sampled at 15 frames per second, the timestamp increment is 6000, based on a 90,000–Hz clock. If an encoded video frame is pack-etized into several RTP packets, each RTP packet will have the same timestamp as the data in each RTP packet are sampled at the same time instant.

A timestamp is particularly useful for media playout control at the receiver. The IP networks usually introduce packet interarrival jitter. In addition, for video encoding, it is possible that a video frame will be skipped in order to maintain a predefined fixed encoding bit rate. By looking at the timestamp, the receiver can properly play out the media at the pace when they were originally sampled. Timestamp information can also be used to synchronize the playout of different media, such as audio and video, with the help of RTCP. We will describe audio and video synchronization in more detail in Section 3.6.3.4.

SSRC specifies the synchronization source and has 32 bits. RTP packets generated from the same source, such as a camera or microphone, should have the same SSRC. SSRC can be used to help the receiver group RTP packets of the same media for playback. CSRC is also

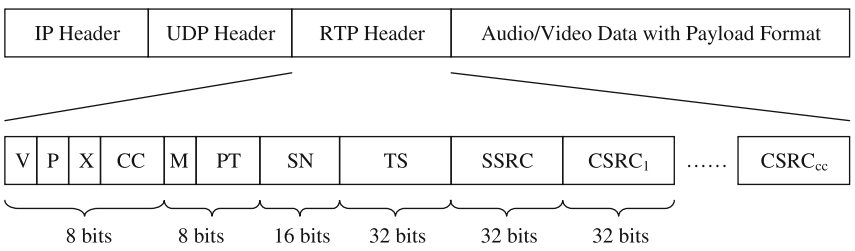


Figure 3.6-5 RTP header format.

a 32-bit field. It indicates the source of a stream of RTP packets that have contributed to the combined stream produced by an RTP mixer. For one-to-one video telephony and VoIP applications, there is no CSRC present in the RTP header.

3.6.2.3.2 Real-time control protocols

RTCP is used in conjunction with RTP to allow RTP session participants to monitor the quality of data delivery. It is based on the periodic transmission of control packets. There are five control packets defined in RFC 3550:

- SR: Send Report. This is sent by an RTP participant that sends and receives the RTP packets;
- RR: Receiver Report. This is sent by an RTP participant that only receives the RTP packets;
- SDES: Source DEScription, including CNAME;
- BYE: This is to indicate the end of the RTP participation;
- APP: Application-specific functions.

Both SR and RR control packets contain reception statistics such as interarrival jitter and packet loss rate. Each SR control packet further includes the sender's wallclock time and the corresponding RTP timestamp when it is generated, as well as transmission statistics, such as how many packets and bytes have been transmitted since the beginning of the RTP session. SR control packets can also be used to synchronize the playout of different media data.

SR and RR reports are also often used for flow and congestion control. For example, by analyzing the interarrival jitter field of the sender report, we can measure the jitter over a certain interval and indicate congestion. As defined in RFC 3550, when packet i is received, the interarrival jitter $J(i)$ is calculated as

$$J(i) = J(i-1) + \frac{|D(i-1,i)| - J(i-1)}{16}, \qquad (3.6.1)$$

where

$$D(i-1,i) = (R(i) - R(i-1)) - (TS(i) - TS(i-1)), \qquad (3.6.2)$$

and $R(i)$ and $TS(i)$ are the arrival time and the timestamp of packet i, respectively. Both are in RTP timestamp units. It is up to the implementation to decide what action to take when congestion occurs. A typical solution is to reduce the transmission rate until congestion becomes alleviated.

The round-trip time can also be estimated using last SR timestamp (LSR) and the delay since last SR (DLSR) information in both RR and SR control packets. Figure 3.6-6 demonstrates one example of a round-trip time calculation. Assume that the RTP sender sends one SR packet at time 10:20:30.250. The RTP receiver receives this SR and, after 5 s, sends an RR packet. In the RR control packet, LSR is the timestamp in SR(i) and the DLSR is 5 s. When the RTP sender receives this RR packet at time 10:20:36.750, it can calculate the round-trip time by subtracting the sending time of SR(i) and the DLSR from the arrival time of RR(i), which is 1.5 s as shown in Figure 3.6-6.

The fraction of loss in SR and RR control packets can also be used for the video encoder to perform error control. The packet loss rate is defined as the number of packets lost over the total number of received packets since the last SR or RR packet was sent.

The transmission interval of RTCP packets is often specified in proportion to the session bandwidth. It is recommended that the fraction of the session bandwidth added for RTCP be fixed at 5%. Some applications may specify the minimal transmission interval to be, for example, 5 s.

3.6.2.3.3 Video payload format

The purposes of using a video payload format are to specify an efficient way to encapsulate data to form a standard-compliant bit stream and to enhance the

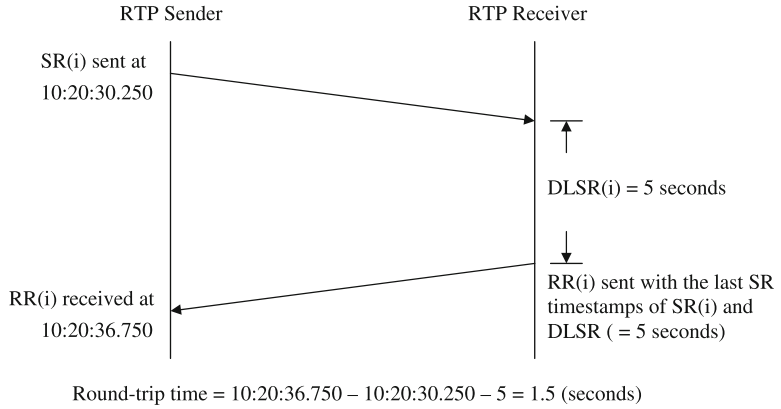


Figure 3.6-6 An example of a round-trip time calculation.

resilience against packet losses. The payload here means media data that are packed in an RTP packet. Forming the media payload can be done in a thin layer between the media encoder and the RTP transport layer. Currently, payload formats defined in RFC 3016 and RFC 2429 for encapsulating MPEG-4 and H.263 video data into individual packets are most commonly used.

The RTP payload formats are designed such that (i) a payload format should be devised so that the stream being transported is still useful even in the presence of a moderate amount of packet loss and (ii) ideally, each packet should possibly be decoded and played out irrespective of whether the preceding packets have been lost or arrived late.

Figure 3.6-7 shows examples of RTP packets generated for MPEG-4 video based on RFC 3016. Among these examples, Figure 3.6-7(b) shows one of the most commonly used packetization methods that have the best error-resilience capability. With this packetization method, one RTP packet contains one video packet. A video packet contains resynchronization marker information at the beginning of the video payload. When the RTP packet containing the VOP header is lost, the other RTP packets can still be decoded due to the use of the Header Extension Code information in the video packet header. No extra RTP header field is necessary.

For H.263 video, similar to MPEG-4 video described in Figure 3.6-7(b), RFC 2429 specifies that the PSC and slice header have to be at the beginning of each RTP packet. It also specifies that the picture header information can be repeated in each RTP packet. This can significantly reduce the number of frames that cannot be decoded due to picture header corruption. H.263 Annex W also provides a similar header protection mechanism, but this repeated header information can only be embedded once in the current picture header or the one in the previous or next frame. Thus, it has lower error resilience and may introduce delay due to waiting for the next frame.

Another purpose of using payload format is for interoperation between two video telephony users that use different applications. A certain video payload format for different codecs has to be supported and implemented to provide a unified video payload encapsulation.

3.6.2.3.4 An example of a call setup process

Before moving to the next section, we provide an example and illustrate the process of setting up a call using SIP. As illustrated in Figure 3.6-8, the caller PC, which is an SIP user agent, initiates a call by sending an INVITE request to the called party. The message has to go through the SIP server that serves the domain of the called party. The SIP server is responsible for locating the addressee via a location service and routing the message to the called party. Once the called party receives the INVITE request, it responses with `2 00 OK`, which is sent back to the caller. Then the caller sends an `ACK` directly to the called party, so that the call is set up, and an RTP pipe is established for audio and video transmission.

SIP, being a signaling protocol, is only responsible for initiating and establishing the session, but the actual communication is directly between the caller and the called party.

3.6.3 Quality of service

3.6.3.1 Minimizing latency

To achieve toll quality for real-time communication, it is typically required that the round-trip delay be lower than 300 ms. Many factors contribute to the packet delay in a real-time communication system. The total end-to-end delay, D, can be divided into the following components:

$$D = d_{\text{enc}} + d_{\text{pack}} + d_{\text{net}} + d_{\text{buf}} + d_{\text{dec},} \tag{3.6.3}$$

where d_{enc} is the encoding delay, d_{pack} is the packetization delay, d_{net} is the delay introduced by the network, d_{buf} is the buffering delay, and d_{dec} is the decoding delay (Figure 3.6-9). To minimize the end-to-end latency, each

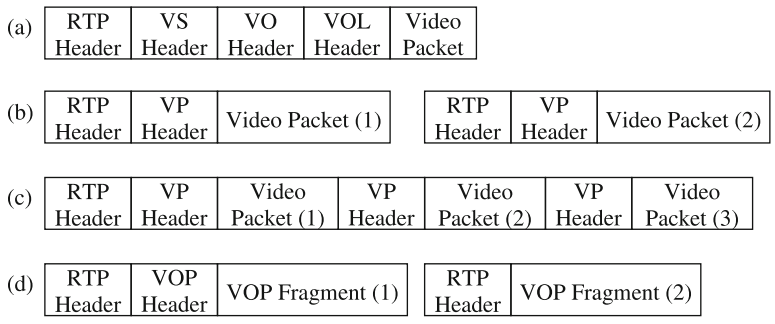


Figure 3.6-7 Examples of MPEG-4 video packetization based on RFC 3016 payload format. VS, visual object sequence; VO, visual object; VOL, visual object layer; VP, video packet; VOP, visual object plane.

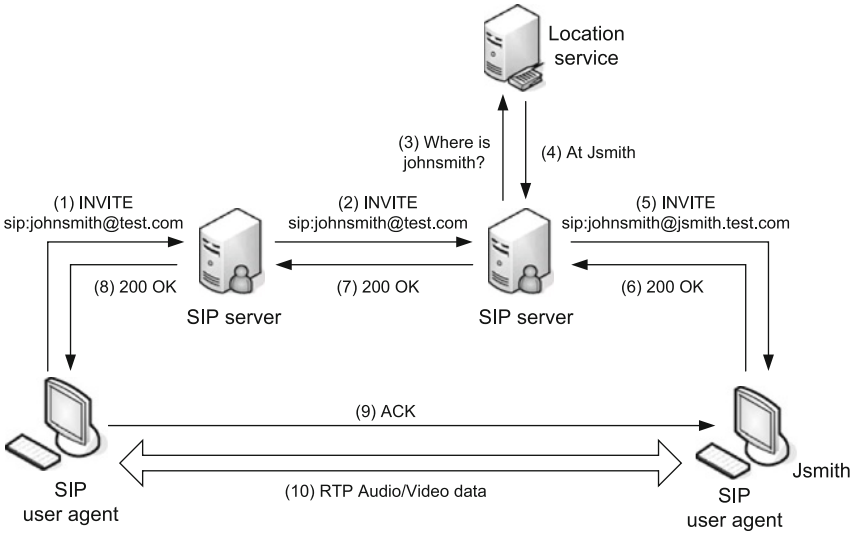


Figure 3.6-8 A call setup process using SIP.

delay component has to be minimized and trade-offs have to be considered in optimizing the overall system design.

Encoding delay is introduced during the data compression process. For speech coders, encoding delay usually includes the frame size and look-ahead delay. Look-ahead delay is the time spent in processing part of the next frame so that a correlation between successive frames can be exploited. Typically, more advanced codecs achieve higher compression efficiency at the cost of higher encoding delays. Decoding delay is introduced during the data decompression process. Table 3.6-3 lists the coding delays for some common speech coders.

Packetization delay is the time spent in collecting sufficient data frames to form the payload of an IP packet. Since the packet headers have a fixed size, a larger payload size reduces the header overhead and improves the transmission efficiency. However, due to the stringent latency requirement, the payload can usually contain only a limited number of frames in order to reduce the packetization delay.

The network delay comprises the propagation delay and the queuing delay across all links in the transmission path. The propagation delay, which is a constant for a fixed path, depends on the packet size and the speed of links, as well as the length of the links. The queuing delay occurs when a packet is queued behind some other packets waiting to be transmitted over the same link. The queuing delay is a random variable depending on the packet size, traffic load and characteristics of the route, and the scheduling scheme. Advanced resource allocation and scheduling schemes such as Resource Reservation Protocol and Differentiated Services enable prioritization of audio and video packets and can efficiently reduce queuing delay for these real-time data streams.

Varying queuing delay, typically caused by congestions of links in the route and related to the queuing mechanisms, introduces delay jitter, which is usually unknown and random. Due to delay jitter, IP packets are sent periodically but are received in irregular patterns. For this reason, a playout buffer, also referred to as a dejitter buffer, is employed at the receiver to absorb the delay jitter before media are output. When using a playout buffer, packets are not played out immediately after being received but are held in a buffer until their scheduled playout time (playout deadline) arrives. Although this introduces additional delay for packets arriving early, this mechanism ensures continuous media playback. The buffering delay is the time a packet is held in the buffer before it is played out.

Note that a trade-off exists between the average buffering delay and the number of packets that have to be dropped because they arrive too late (late loss). Scheduling a later deadline increases the possibility of playing out more packets and results in a lower loss rate, but at the

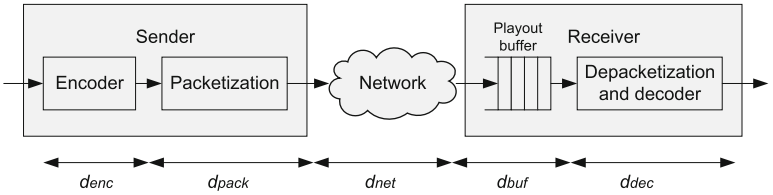


Figure 3.6-9 Total end-to-end delay in a typical real-time communication system.

Table 3.6-3 Coding delays for sample speech coders.

Speech coder	Encoded bit rate (kbit/s)	Frame size (ms)	Look-ahead delay (ms)	Decoding delay (ms)
G.711	64	0.125	0	0
G.729A	8	10	5	7.5
G.723.1	5.3/6.4	30	7.5	18.75

cost of a higher buffering delay. Vice versa, it is difficult to decrease the buffering delay without significantly increasing the loss rate. Therefore, packet loss in delay-sensitive, real-time applications is a result of not only a packet being dropped over the network, but also delay jitter, which impairs communication quality greatly.

Due to the aforementioned buffering delay—late loss rate trade-off—it is desirable to design smart playout scheduling mechanisms to reduce the buffering delay. Fixed scheduling poses a limitation for this trade-off. In real-time speech communication, more advanced mechanisms use a playout buffer to completely absorb delay jitter within talkspurts and dynamically adjust the schedule between talkspurts [12–16]. Adaptive playout scheduling is proposed to allow adaptive schedules even within talkspurts [17], and this idea has also been extended to video streaming [18,19]. An adaptive play out schedule is able to reduce the latency and the effective packet loss rate at the same time.

3.6.3.2 Combating losses

In real-time communications, losses are a result of not only packets dropping over the network, but also late arrival for packets. We introduce different loss-resilient techniques for both audio and video in two categories: client-side techniques and active techniques, depending on whether they require any encoder involvement.

3.6.3.2.1 Client-side techniques

One category of loss techniques is passive methods that are implemented at the client side, which do not require any cooperation of the sender or increase the cost of transmission. Client-side techniques impose low overhead for the communication system but can be highly efficient in enhancing the quality of the rendered media.

To combat channel losses, the client typically employs error-detection and loss-concealment techniques to mitigate the effect of lost data. For speech and audio coded techniques based on waveform, most client-side schemes take advantage of the data received adjacent to the lost packet and interpolate the missing information by exploiting the redundancy in the signal. In particular, waveform repetition simply repeats the information contained in the packets prior to the lost one [20, 21]. A more advanced loss-concealment technique using time-scale modification is described in [22] and [23] and can be used in conjunction with adaptive playout scheduling in a low-latency scenario [17]. Waveform repetition typically does not introduce any algorithm delay as timescale modification typically does. However, it does not provide as good a sound quality [24].

For video communication, postprocessing is typically applied at the client side for error concealment and loss recovery. Techniques to recover the damaged areas based on characteristics of image and video signals have been reviewed in [25].

3.6.3.2.2 Active techniques

A different category of error-resilience techniques requires the encoder to play a primary role. They are able to provide even higher robustness for media communication over best-effort networks. We refer to these techniques as "active" to differentiate them from those only employed at the client side.

For speech communication, one widely accepted way to reduce the effective packet loss observed by the receiver is to add redundancy to the data stream at the sender. This is possible without imposing too much extra network load since the data rate of the voice traffic is very low when compared with other types of multimedia and data traffic. A common method to add redundancy is forward error correction (FEC), which transmits redundant information across packets, where loss recovery is performed at the cost of higher latency. The efficiency of FEC schemes is largely limited by the bursty nature of the channel losses. In order to combat burst losses, redundant information has to be added into temporally distant packets, which introduce higher delay.

Another sender-based loss recovery technique, interleaving, does not increase the data rate of transmission but still introduces delay at both encoder and decoder sides. The efficiency of loss recovery depends on over how many packets the source packet is interleaved and spread over. Again, the wider the spread, the higher the introduced delay. For low-latency speech communication, path diversity techniques, presented in [26], have been demonstrated to be very powerful in combating losses.

Video communication typically requires much higher data transmission rates than audio. A variety of active schemes has been proposed not only to increase the robustness of communication, but also to take the data rate efficiency into consideration [27–29]. Many of the recent algorithms use rate–distortion (R–D) optimization techniques to improve the compression efficiency [30–32], as well as to improve the error-resilient performance

over lossy networks [33,34]. The goal of the R–D optimization algorithms is to minimize the expected distortion due to both compression and channel losses subject to the bit-rate constraint.

One example of this area is Intra/Inter-mode switching [35–38], where Intra-coded macroblocks are updated according to the network condition to mitigate temporal error propagation. Another approach is to modify the temporal prediction dependency of motion-compensated video coding in order to mitigate or stop error propagation. Example implementations include reference picture selection [27,39–41] and NEWPRED in MPEG-4 [42,43], where channel feedback is used to efficiently stop error propagation due to any transmission error. Another example is video redundancy coding (VRC), where the video sequence is coded into independent threads (streams) in a round-robin fashion [27,44]. A Sync-frame is encoded by all threads at regular intervals to start a new thread series and stop error propagation. If one thread is damaged due to packet loss, the remaining threads can still be used to predict the Sync-frame. VRC provides improved error resilience, but at the cost of a much higher data rate. Dynamic control of the prediction dependency can also be used by employing long-term memory prediction to achieve improved R–D performance [33,45,46].

Typically a channel coding module in a robust video communication system may involve FEC and automatic retransmission on request (ARQ). Similar to their applications in speech communication, when FEC is employed across packets, missing packets can be recovered at the receiver as long as a sufficient number of packets is received [47–50]. FEC is widely used as an unequal error protection scheme to protect prioritized transmissions. In addition to FEC codes, data randomization and interleaving are also employed for enhanced protection [51–55].

ARQ techniques incorporate channel feedback and employ the retransmission of erroneous data [56–60]. Unlike FEC schemes, ARQ intrinsically adapts to the varying channel conditions and tends to be more efficient in transmission. However, for real-time communication and low-latency streaming, the latency introduced by ARQ is a major concern. In addition, like all feedback-based error control schemes, ARQ is not appropriate for multicasting.

3.6.3.3 Adapting to the available bandwidth

Due to the lack of a QoS guarantee over most commercially deployed networks, it is expected that the condition, as well as the available bandwidth of the network, varies during a real-time communication session. It is beneficial to employ bandwidth adaptation mechanisms to control the rate at which the media are transmitted. This helps avoid a potential penalty on overuse of bandwidth, which usually leads to quality degradation and even service interruption. Typical bandwidth adaptation techniques include rate control, transcoding, scalable coding, and bit stream switching.

3.6.3.4 Audio–video synchronization

RTP timestamps from different media streams may advance at different rates and usually have independent and random offsets. Therefore, although these timestamps are sufficient to reconstruct the timing of a single stream, directly comparing RTP timestamps from different media is not effective for synchronization. Instead, for each medium the RTP timestamp is related to the sampling instant by pairing it with a timestamp from a reference clock (wallclock) that represents the time when the medium was sampled. The reference clock is shared by all media to be synchronized.

Synchronizing audio and video can be achieved by playing out audio and video according to their original sampled time. By doing so, the receiver can play back audio and video at a proper pace by mapping their original sampled time to the receiver's local time. RTCP SR control packets provide useful information to help the receiver calculate the sampled time of the audio and the video at the sender. Figure 3.6-10 illustrates an example of audio and video synchronization by using RTCP SR control packets. When an RTCP SR control packet is generated, it will carry the wallclock time (NTP) and the RTP timestamp using its corresponding media reference time. In Figure 3.6-10, the RTCP SR control packet for the audio RTP session is generated at time 10:20:30.730 and the corresponding timestamp is 200. When receiving this SR packet, the receiver is able to calculate when all the audio RTP packets are sampled at the sender. For the example in Figure 3.6-10, the audio packet with timestamp 160 is actually generated at time 10:20:30.725, assuming that a 8000-Hz clock is used for audio timestamping. Similarly, for video packets, the receiver can also calculate when each video frame is sampled. In this way, the receiver can easily find out which part of audio data and video data should be played back at the same time.

3.6.4 Summary and further reading

In this chapter, we have described the system and architecture for real-time communication, including two categories of the most important protocols, signaling and transport, respectively. We have also addressed the QoS issues, especially on minimizing latency, combating losses, adapting to available bandwidth, and audio-video synchronization.

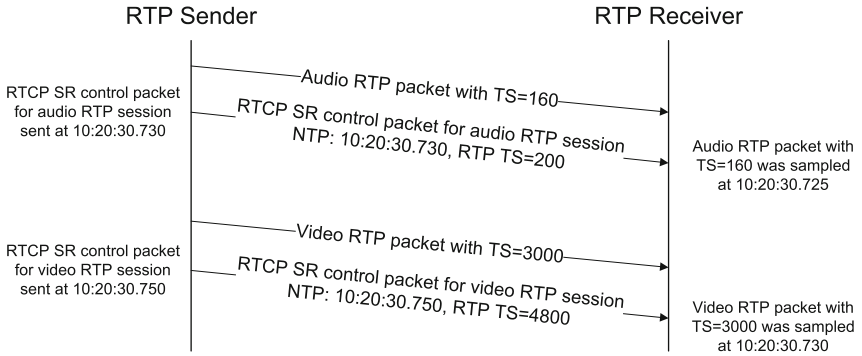


Figure 3.6-10 An example of audio and video synchronization.

References

[1] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, M. Hardley, and E. Schooler. "SIP: Session Initiation Protocol," *RFC 3261*, June 2002.

[2] S.Donovan. "The SIP INFO Method," *RFC 2976*, Oct. 2000.

[3] B. Campbell et al. "Session Initiation Protocol (SIP) Extension for Instant Messaging," *RFC 3428*, December 2002.

[4] A.B. Roach. "Session Initiation Protocol (SIP): Specific Event Notification," *RFC 3265*, June 2002.

[5] J. Rosenberg and H. Schulzrinne. "Reliability of Provisional Responses in the Session Initiation Protocol," *RFC 3262*, June 2002.

[6] R. Sparks. "The Session Initiation Protocol (SIP) Refer Method," *RFC 3515*, April 2003.

[7] J. Rosenberg. "The Session Initiation Protocol (SIP) UPDATE Method," *RFC 3311*, September 2002.

[8] M. Handley and V. Jacobson. "SDP: Session Description Protocol," *RFC 2327*, April 1998.

[9] S. Sjoberg, M. Westerlurd, A. Lakaniemi, and Q. Xie. "Real-Time Transport Protocol (RTP) Payload Format and File Storage Format for the Adaptive Multi-Rate (AMR) and Adaptive Multi-Rate Wideband (AMR-WB) Audio Codec," *RFC 3267*, June 2002.

[10] Y. Kikuchi, T. Nomura, S. Fukunaga, Y. Matsui, and H. Kimata. "RTP Payload Format for MPEG-4 Audio/ Visual Streams," *RFC 3016*, November 2000.

[11] I. Dalgic and H. Fang. "Comparison of H.323 and SIP for IP Telephony Signaling," in *Proc. of Photonics East*, Boston, MA, September 1999.

[12] R. Ramjee, J. Kurose, D. Towsley, and H. Schulzrinne. "Adaptive Playout Mechanisms for Packetized Audio Applications in Wide-Area Networks," in *Proceedings IEEE INFOCOM '94*, vol. 2, pp. 680–688, June 1994.

[13] S.B. Moon, J. Kurose, and D. Towsley. "Packet Audio Playout Delay Adjustment: Performance Bounds and Algorithms," *Multimedia Systems*, vol. 6, no. 1, pp. 17–28, January 1998.

[14] J. Pinto and K.J. Christensen. "An Algorithm for Playout of Packet Voice Based on Adaptive Adjustment of Talkspurt Silence Periods," in *Proceedings 24th Conference on Local Computer Networks*, pp. 224–231, October 1999.

[15] P. DeLeon and C.J. Sreenan. "An Adaptive Predictor for Media Playout Buffering," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-99)*, vol. 6, pp. 3097–3100, March 1999.

[16] J. Rosenberg, L. Qiu, and H. Schulzrinne. "Integrating Packet FEC into Adaptive Voice Playout Buffer Algorithms on the Internet," in *Proceedings IEEE INFOCOM 2000*, vol. 3, pp. 1705–1714, Tel Aviv, Israel, March 2000.

[17] Y.J. Liang, N. Färber, and B. Girod. "Adaptive Playout Scheduling and Loss Concealment for Voice Communication over IP Networks," *IEEE Transactions on Multimedia*, vol. 5, no. 4, pp. 532–543, December 2003.

[18] E. Steinbach, N. Färber, and B. Girod. "Adaptive Playout for Low-Latency Video Streaming," in *IEEE International Conference on Image Processsing ICIP-01*, vol. 1, pp. 962–965, Thessaloniki, Greece, October 2001.

[19] M. Kalman, E. Steinbach, and B. Girod. "Adaptive Playout for Real-Time Media Streaming," in *IEEE International Symposium on Circuits and Systems*, vol. 1, pp. I–45–8, Scottsdale, AZ, May 2002.

[20] D.J. Goodman, G.B. Lockhart, O.J. Wasem, and W.-C. Wong. "Waveform Substitution Techniques for Recovering Missing Speech Segments in Packet Voice Communications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, no. 6, pp. 1440–1448, December 1986.

[21] O.J. Wasem, D.J. Goodman, C.A. Dvorak, and H.G. Page. "The Effect of Waveform Substitution on the Quality of PCM Packet Communications," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 3, pp. 342–348, March 1988.

[22] A. Stenger, K. Ben Younes, R. Reng, and B. Girod. "A New Error Concealment Technique for Audio Transmission with Packet Loss," in *Proc. European Signal Processing Conference*, vol. 3, pp. 1965–1968, September 1996.

[23] H. Sanneck, A. Stenger, K. Ben Younes, and B. Girod. "A New Technique for Audio Packet Loss Concealment," in *IEEE GLOBECOM*, pp. 48–52, November 1996.

[24] C. Perkins, O. Hodson, and V. Hardman. "A Survey of Packet Loss Recovery Techniques for Streaming Audio," *IEEE Network*, vol. 12, no. 5, pp. 40–48, September–October 1998.

[25] Yao Wang and Qin-Fan Zhu. "Error Control and Concealment for Video Communication: A Review," *Proceedings of the IEEE*, vol. 86, no. 5, pp. 974–997, May 1998.

[26] Y.J. Liang, E.G. Steinbach, and B. Girod. "Real-Time Voice Communication over the Internet Using Packet Path Diversity," in *Proceedings ACM Multimedia 2001*, pp. 431–440, Ottawa, Canada, October 2001.

[27] S. Wenger, G.D. Knorr, J. Ott, and F. Kossentini. "Error Resilience Support in H.263+," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 8, no. 7, pp. 867–877, November 1998.

[28] R. Talluri. "Error-Resilient Video Coding in the ISO MPEG-4 Standard," *IEEE Communications Magazine*, pp. 112–119, June 1998.

[29] W. Tan and A. Zakhor. "Real-time Internet Video Using Error Resilient Scalable Compression and TCP-Friendly Transport Protocol," *IEEE Trans. Multimedia*, pp. 172–186, June 1999.

[30] G.J. Sullivan and T. Wiegand. "Rate-Distortion Optimization for Video Compression," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 74–90, November 1998.

[31] A. Ortega and K. Ramchandran. "From Rate-Distortion Theory to Commercial Image and Video Compression Technology," *IEEE Signal Processing Magazine*, vol. 15, no. 6, pp. 20–122, November 1998.

[32] T. Wiegand, X. Zhang, and B. Girod. "Long-Term Memory Motion-Compensated Prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, February 1999.

[33] T. Wiegand, N. Färber, and B. Girod. "Error-Resilient Video Transmission Using Long-Term Memory Motion-Compensated Prediction," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1050–1062, June 2000.

[34] P.A. Chou, A.E. Mohr, A. Wang, and S. Mehrotra, "Error Control for Receiver-Driven Layered Multicast of Audio and Video," *IEEE Transactions on Multimedia*, vol. 3, no. 1, pp. 108–122, March 2001.

[35] J.Y. Liao and J.D. Villasenor. "Adaptive Intra Update for Video Coding over Noisy Channels," in *Proceedings IEEE International Conference on Image Processing*, Lausanne, Switzerland, vol. 3, pp. 763–766, September 1996.

[36] R.O. Hinds, T.N. Pappas, and J.S. Lim. "Joint Block-Based Video Source/Channel Coding for Packet-Switched Networks," in *Proceedings of the SPIE VCIP 98*, vol. 3309, pp. 124–133, San Jose, CA, October 1998.

[37] G. Cote and F. Kossentini. "Optimal Intra Coding of Blocks for Robust Video Communication over the Internet," *Signal Processing: Image Communication*, vol. 15, no. 1-2, pp. 25–34, September 1999.

[38] R. Zhang, S.L. Regunathan, and K. Rose. "Video Coding with Optimal Inter/Intra-mode Switching for Packet Loss Resilience," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 966–976, June 2000.

[39] S. Fukunaga, T. Nakai, and H. Inoue. "Error Resilient Video Coding by Dynamic Replacing of Reference Pictures," in *Proc. of the IEEE Global Telecommunications Conference*, vol. 3, pp. 1503–1508, London, UK, November 1996.

[40] ITU-T Recommendation H.263 Version 2 (H.263+), *Video Coding for Low Bitrate Communication*, January 1998.

[41] ITU-T Recommendation H.264, *Advanced Video Coding (AVC)for Generic Audiovisual Services*, May 2003.

[42] International Organisation for Standardisation, ISO/IEC JTC1/ SC29/WG11 Final Committee Draft 14496-2, *Information Technology – Coding of Audio-Visual Objects: Visual (MPEG-4)*, March 1998.

[43] International Organisation for Standardisation, ISO/IEC JTC1/ SC29/WG11 Final Committee Draft 14496-2, *Information Technology – Coding of Audio-Visual Objects: Visual (MPEG-4)*, March 1998.

[44] S. Wenger. "Video Redundancy Coding in H.263+," in *Proc. of the Workshop on Audio-Visual Services for Packet Networks*, September 1997.

[45] M. Budagavi and J.D Gibson. "Multiframe Video Coding for Improved Performance over Wireless Channels," *IEEE Transactions on Image Processing*, vol. 10, no. 2, pp. 252–265, February 2001.

[46] Y.J. Liang and B. Girod. "Network-Adaptive Low-Latency Video Communication over Best-Effort Networks," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 1, pp. 72–81, January 2006.

[47] Internet Engineering Task Force. "RTP Pay load Format for MPEG-1 /MPEG-2 Video," *RFC 2250*, January 1998.

[48] A. Albanese, J. Blömer, J. Edmonds, M. Luby, and M. Sudan. "Priority Encoding Transmission," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 1737–1744, November 1996.

[49] P.C. Cosman, J.K. Rogers, P.G. Sherwood, and K. Zeger. "Image Transmission over Channels with Bit Errors and Packet Erasures," in *Proceedings of the Thirty-Second Asilomar Conference on Signals, Systems and Computers*, vol. 2, pp. 1621–1625, Pacific Grove, CA, November 1998.

[50] W. Tan and A. Zakhor. "Video Multicast Using Layered Fec and Scalable Compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 3, pp. 373–387, March 2001.

[51] J.-Y. Cochennec. *Method for the Correction of Cell Losses for Low Bit-Rate Signals Transport with the AAL Type 1*. ITU-T SG15 Doc. AVC-538, July 1993.

[52] Q.-F. Zhu, Y. Wang, and L. Shaw. "Coding and Cell-Loss Recovery in DCT-Based Packet Video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, no. 3, pp. 248–258, June 1993.

[53] T. Kinoshita, T. Nakahashi, and M. Maruyama, "Variable-Bit-Rate HDTV Codec with ATM-Cell-Loss Compensation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 3, no. 3, pp. 230–237, June 1993.

[54] K. Stuhlmüller, N. Färber, M. Link, and B. Girod. "Analysis of Video

Transmission over Lossy Channels," *IEEE Journal on Selected Areas in Communications*, vol. 18, no. 6, pp. 1012–1032, June 2000.

[55] Y.J. Liang, J.G. Apostolopoulos, and B. Girod, "Model-Based Delay-Distortion Optimization for Video Streaming Using Packet Interleaving," in *Proceedings of the 36th Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, November 2002.

[56] S.B. Wicker. *Error Control Systems for Digital Communication and Storage*, Prentice Hall, 1995.

[57] M. Khansari, A. Jalali, E. Dubois, and P. Mermelstein. "Low Bit-Rate Video Transmission over Fading Channels for Wireless Microcellular Systems," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 6, no. 1, pp. 1–11, February 1996.

[58] B. Dempsey, J. Liebeherr, and A. Weaver. "On Retransmission-Based Error Control for Continuous Media Traffic in Packet-Switching Networks," *Computer Networks and ISDN Systems Journal*, vol. 28, no. 5, pp. 719–736, March 1996.

[59] C. Papadopoulos and G.M. Parulkar. "Retransmission-Based Error Control for Continuous Media Applications," in *Proc. Network and Operating System Support for Digital Audio and Video (NOSSDAV)*, Zushi, Japan, July 1996.

[60] H. Liu and M. El Zarki. "Performance of H.263 Video Transmission over Wireless Channels Using Hybrid ARQ," *IEEE Journal on Selected Areas in Communications*, vol. 15, no. 9, pp. 1775–1786, December 1999.

This page intentionally left blank

Chapter 3.7

3.7

Wireless sensor networks

Feng Zhao and Leonidas Guibas

The most profound technologies are those that disappear. They weave themselves into the fabric of everyday life until they are indistinguishable from it.

—The late Mark Weiser, Father of Ubiquitous Computing and Chief Technologist of Xerox PARC

Advances in wireless networking, micro-fabrication and integration (for example, sensors and actuators manufactured using micro-electromechanical system technology, or MEMS), and embedded microprocessors have enabled a new generation of massive-scale sensor networks suitable for a range of commercial and military applications. The technology promises to revolutionize the way we live, work, and interact with the physical environment. In the not-too-distant future, tiny, dirt-cheap sensors may be literally sprayed onto roads, walls, or machines, creating a digital skin that senses a variety of physical phenomena of interest: monitor pedestrian or vehicular traffic in human-aware environments and intelligent transportation grids, report wildlife habitat conditions for environmental conservation, detect forest fires to aid rapid emergency response, and track job flows and supply chains in smart factories. Unlike current information services such as those on the Internet where information can easily get stale or be useless because it is too generic, sensor networks promise to couple end users directly to sensor measurements and provide information that is precisely localized in time and/or space, according to the user's needs or demands.

With such technological advances come new challenges for information processing in sensor networks. What is needed are novel computational representations, algorithms and protocols, and design methodologies and tools to support distributed signal processing, information storage and management, networking, and application development. While this book will primarily focus on wireless sensor networks, some of the principles, such as those of collaborative information processing and management, apply equally well to wireline sensor networks. The issues of scalability and efficient use of bandwidth, a main concern of the book, are common to both wireless and wireline sensor networks.

3.7.1 Unique constraints and challenges

Unlike a centralized system, a sensor network is subject to a unique set of resource constraints such as finite on-board battery power and limited network communication bandwidth. In a typical sensor network, each sensor node operates untethered and has a microprocessor and a small amount of memory for signal processing and task scheduling. Each node is also equipped with one or more sensing devices such as acoustic microphone arrays, video or still cameras, infrared (IR), seismic, or magnetic sensors. Each sensor node communicates wirelessly with a few other local nodes within its radio communication range.

Sensor networks extend the existing Internet deep into the physical environment. The resulting new network is orders of magnitude more expansive and dynamic than the current TCP/IP network and is creating entirely new types of traffic that are quite different from what one finds on the Internet now. Information collected by and transmitted on a sensor network describes conditions of physical environments—for example, temperature, humidity, or vibration—and requires advanced query interfaces and search engines to effectively support

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2004 Elsevier Inc. All rights reserved

user-level functions. Sensor networks may internetwork with an IP core network via a number of gateways, as in Figure 3.7-1. A gateway routes user queries or commands to appropriate nodes in a sensor network. It also routes sensor data, at times aggregated and summarized, to users who have requested it or are expected to utilize the information. A data repository or storage service may be present at the gateway, in addition to data logging at each sensor.

The repository may serve as an intermediary between users and sensors, providing a persistent data storage. Additionally, one or more data storage devices may be attached to the IP network, to archive sensor data from a number of edge sensor networks and to support a variety of user-initiated browsing and search functions.

The current generation of wireless sensor hardware ranges from shoe-box-sized Sensoria WINS NG sensors with an SH-4 microprocessor to matchbox-sized Berkeley motes with an 8-bit microcontroller. A few samples of sensor hardware are shown in Figure 3.7-2; their corresponding capabilities are summarized and compared in Table 3.7-1. It is well known that communicating 1 bit over the wireless medium at short ranges consumes far more energy than processing that bit. For the Sensoria sensors and Berkeley motes, the ratio of energy consumption for communication and computation is in the range of 1000 to 10,000. Despite the advances in silicon fabrication technologies, wireless communication will continue to dominate the energy consumption of networked embedded systems for the foreseeable future.

Thus, minimizing the amount and range of communication as much as possible—for example, through local collaboration among sensors, duplicate data suppression, or invoking only the nodes that are relevant to a given task—can significantly prolong the life of a sensor network and leave nodes free to support multiuser operations. In addition, the shorter RF transmission range improves spectrum usage and increases throughput for a sensor network.

The information management and networking for this new network will require more than just building faster routers, switchers, and browsers. A sensor network is designed to collect information from a physical environment. Networking will be intimately coupled with the needs of sensing and control, and hence the application semantics. To optimize for performance and resources such as energy, one has to rethink the existing TCP/IP stack and design an appropriate sensor network abstraction to support application development. For example, in many applications, it is more appropriate to

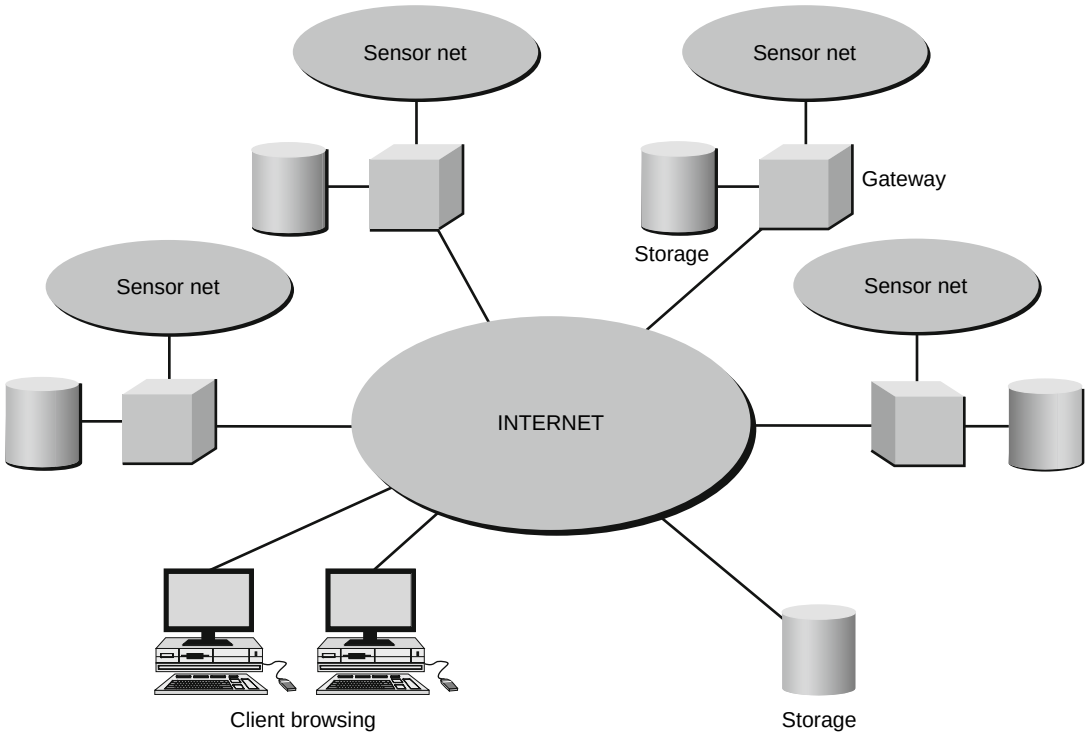


Figure 3.7-1 Sensor networks significantly expand the existing Internet into physical spaces. The data processing, storage, transport, querying, as well as the internetworking between the TCP/IP and sensor networks present a number of interesting research challenges that must be addressed from a multidisciplinary, cross-layer perspective.

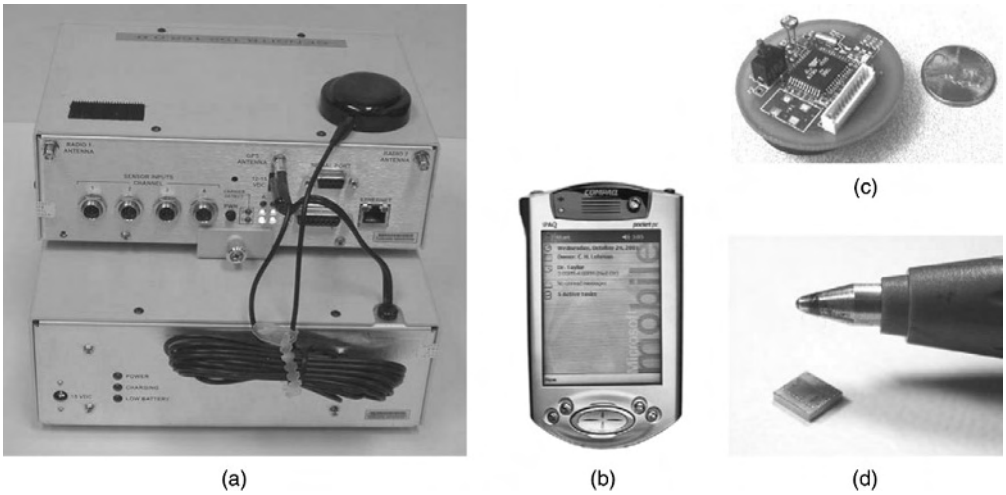


Figure 3.7-2 Samples of wireless sensor hardware: (a) Sensoria WINS NG 2.0 sensor node; (b) HP iPAQ with 802.11b and microphone; (c) Berkeley/Crossbow sensor mote, alongside a U.S. penny; (d) An early prototype of Smart Dust MEMS integrated sensor, being developed at UC Berkeley. (*Picture courtesy of Kris Pister and Jason Hill*).

address nodes in a sensor network by physical properties, such as node locations or proximity, than by IP addresses. How and where data is generated by sensors and consumed by users will affect the way data is compressed, routed, and aggregated. Because of the peer-to-peer connectivity and the lack of a global infrastructure support, the sensors have to rely on discovery protocols to construct local models about the network and environment. Mobility and instability in wireless links preclude the use of many existing edge-network gateway protocols for internetworking IP and sensor networks.

Table 3.7-1 Comparison of the four sensor platforms shown in Figure 3.7-2.

	WINS NG 2.0 Node	iPAQ with 802.11 and A/D Cards in Sleeve	Berkeley MICA Mote*	Smart Dust**
Parts cost*** (quantity 1000+)	$100s	$100s	$10s	<$1
Size (cm^3)	5300	600	40	.002
Weight (*g*) (including battery)	5400	350	70	.002
Battery capacity *(kJ)*	300	35	15	(Less)
Sensors	Off-board	Microphone & light sensors integrated, others off-board	Integrated on PCB: Acceleration, temperature, light, sound	MEMS sensors to be integrated
Memory	32 MB RAM, 32 MB flash	64 MB RAM, 32 MB flash	4 KB RAM, 128 KB flash	(Less)
CPU	Hitachi SH4	StrongARM or XScale	ATmega 103L	(Less powerful)
Operating system	Linux	WinCE or Linux	TinyOS	(smaller)
Processing capability	400 MIPS/1.4 GFLOPS	240 MIPS	4 MIPS	(Less)
Radio range	100 m	100 m	30 m	(Shorter)

* The MICA mote is slightly larger than the WeC mote shown in Figure 3.7-2(c), and is more widely used.
** Smart Dust is not yet fully operational, but the size goal and power sources are known, and cost and weight are estimated.
*** Note that the parts cost is based on large-quantity production.

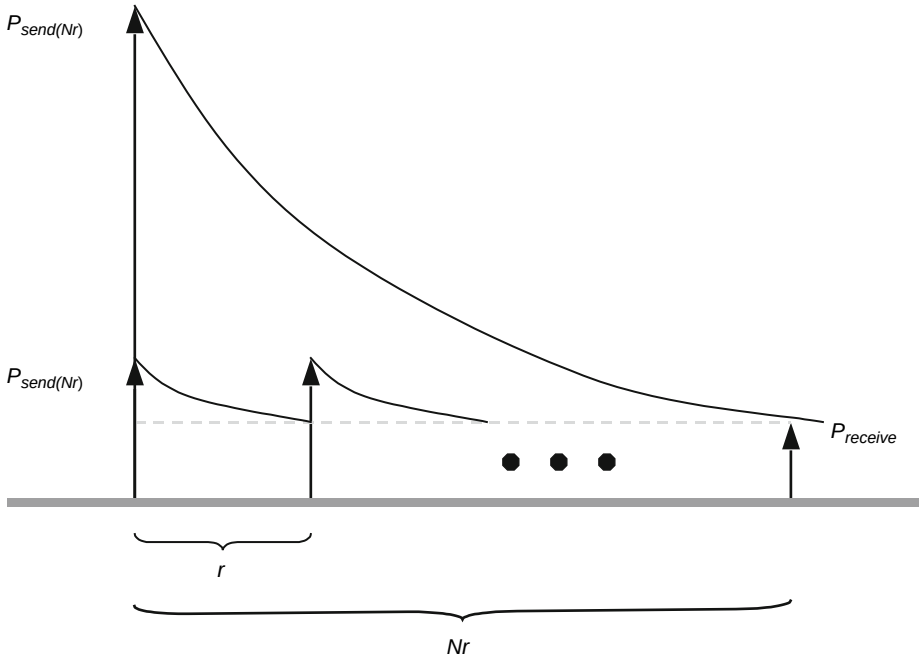


Figure 3.7-3 The power advantage of using a multihop RF communication over a distance of *Nr*.

To summarize, the challenges we face in designing sensor network systems and applications include:

- *Limited hardware:* Each node has limited processing, storage, and communication capabilities, and limited energy supply and bandwidth.
- *Limited support for networking:* The network is peer-to-peer, with a mesh topology and dynamic, mobile, and unreliable connectivity. There are no universal routing protocols or central registry services. Each node acts both as a router and as an application host.
- *Limited support for software development:* The tasks are typically real-time and massively distributed, involve dynamic collaboration among nodes, and must handle multiple competing events. Global properties can be specified only via local instructions. Because of the coupling between applications and system layers, the software architecture must be codesigned with the information processing architecture.

3.7.2 Advantages of sensor networks

Networked sensing offers unique advantages over traditional centralized approaches. Dense networks of distributed communicating sensors can improve signal-to-noise ratio (SNR) by reducing average distances from sensor to source of signal, or target. Increased energy efficiency in communications is enabled by the multihop topology of the network. Moreover, additional relevant information from other sensors can be aggregated during this multihop transmission through in-network processing. But perhaps the greatest advantages of networked sensing are in improved robustness and scalability. A decentralized sensing system is inherently more robust against individual sensor node or link failures, because of redundancy in the network. Decentralized algorithms are also far more scalable in practical deployment and may be the only way to achieve the large scales needed for some applications.

3.7.2.1 Energy advantage

Because of the unique attenuation characteristics of radio-frequency (RF) signals, a multihop RF network provides a significant energy saving over a single-hop network for the same distance. Consider the following simple example of an *N*-hop network. Assume the overall distance for transmission is *Nr*, where *r* is the one-hop distance. The minimum receiving power at a node for a given transmission error rate is $P_{receive}$, and the power at a transmission node is P_{send}. Then, the RF attenuation model near the ground is given by

$$P_{receive} \propto \frac{P_{send}}{r^{\alpha}},$$

where *r* is the transmission distance and α is the RF attenuation exponent. Due to multipath and other

interference effects, α is typically in the range of 2 to 5. Equivalently,

$$P_{send} \propto r^{\alpha} P_{receive}.$$

Therefore, the power advantage of an N-hop transmission versus a single-hop transmission over the same distance Nr is

$$\eta_{rf} = \frac{P_{send(Nr)}}{N \cdot P_{send(r)}} = \frac{(Nr)^{\alpha} P_{receive}}{N \cdot r^{\alpha} P_{receive}} = N^{\alpha-1}. \quad (3.7.1)$$

Figure 3.7-3 illustrates the power attenuation for the multihop and single-hop networks. A larger N gives a larger power saving due to the consideration of RF energy alone. However, this analysis ignores the power usage by other components of an RF circuitry. Using more nodes increases not only the cost, but also the power consumption of these other RF components. In practice, an optimal design seeks to balance the two conflicting factors for an overall cost and energy efficiency. Latency and robustness considerations may also argue against an unduly large number of relay nodes.

3.7.2.2 Detection advantage

Each sensor has a finite sensing range, determined by the noise floor at the sensor. A denser sensor field improves the odds of detecting a signal source within the range. Once a signal source is inside the sensing range of a sensor, further increasing the sensor density decreases the average distance from a sensor to the signal source, hence improving the signal-to-noise ratio (SNR). Let us consider the acoustic sensing case in a two-dimensional plane, where the acoustic power received at a distance r is

$$P_{receive} \propto \frac{P_{source}}{r^2},$$

which assumes an inverse distance squared attenuation. The SNR is given by

$$\begin{aligned} \mathrm{SNR}_r &= 10 \log \frac{P_{receive}}{P_{noise}} \\ &= 10 \log P_{source} - 10 \log P_{noise} - 20 \log r. \end{aligned}$$

Increasing the sensor density by a factor of k reduces the average distance to a target by a factor of $\frac{1}{\sqrt{k}}$. Thus, the SNR advantage of the denser sensor network is

$$\eta_{snr} = \mathrm{SNR}_{\frac{r}{\sqrt{k}}} - \mathrm{SNR}_r = 20 \log \frac{r}{\frac{r}{\sqrt{k}}} = 10 \log k. \quad (3.7.2)$$

Therefore, an increase in sensor density by a factor of k improves the SNR at a sensor by 10 log k db.

3.7.3 Sensor network applications

A sensor network is designed to perform a set of high-level information processing tasks such as detection, tracking, or classification. Measures of performance for these tasks are well defined, including detection of false alarms or misses, classification errors, and track quality. Applications of sensor networks are wide ranging and can vary significantly in application requirements, modes of deployment (e.g., ad hoc versus instrumented environment), sensing modality, or means of power supply (e.g., battery versus wall-socket). Sample commercial and military applications include:

- Environmental monitoring (e.g., traffic, habitat, security)
- Industrial sensing and diagnostics (e.g., appliances, factory, supply chains)
- Infrastructure protection (e.g., power grids, water distribution)
- Battlefield awareness (e.g., multitarget tracking)
- Context-aware computing (e.g., intelligent home, responsive environment)

Now we look only at a few application samples in order to ground the technical discussions for the rest of the book.

3.7.3.1 Habitat monitoring: Wildlife conservation through autonomous, nonintrusive sensing

On a small patch of land 10 miles off the coast of Maine, a team of computer engineers from the University of California, Berkeley, are conducting an experiment in networked sensing. Working with biologists at the College of the Atlantic, the engineers have installed 190 wireless sensors that are being used to monitor the habitat of the nesting petrels[1] on Great Duck Island (see Figure 3.7-4). In the past, biologists studying the nesting behaviors of these birds had to travel to the island every now and then to gather observation data. To check on the

[1] This is the name given to various ocean birds belonging to the order of tube-nosed swimmers, like the albatross and the shearwater.

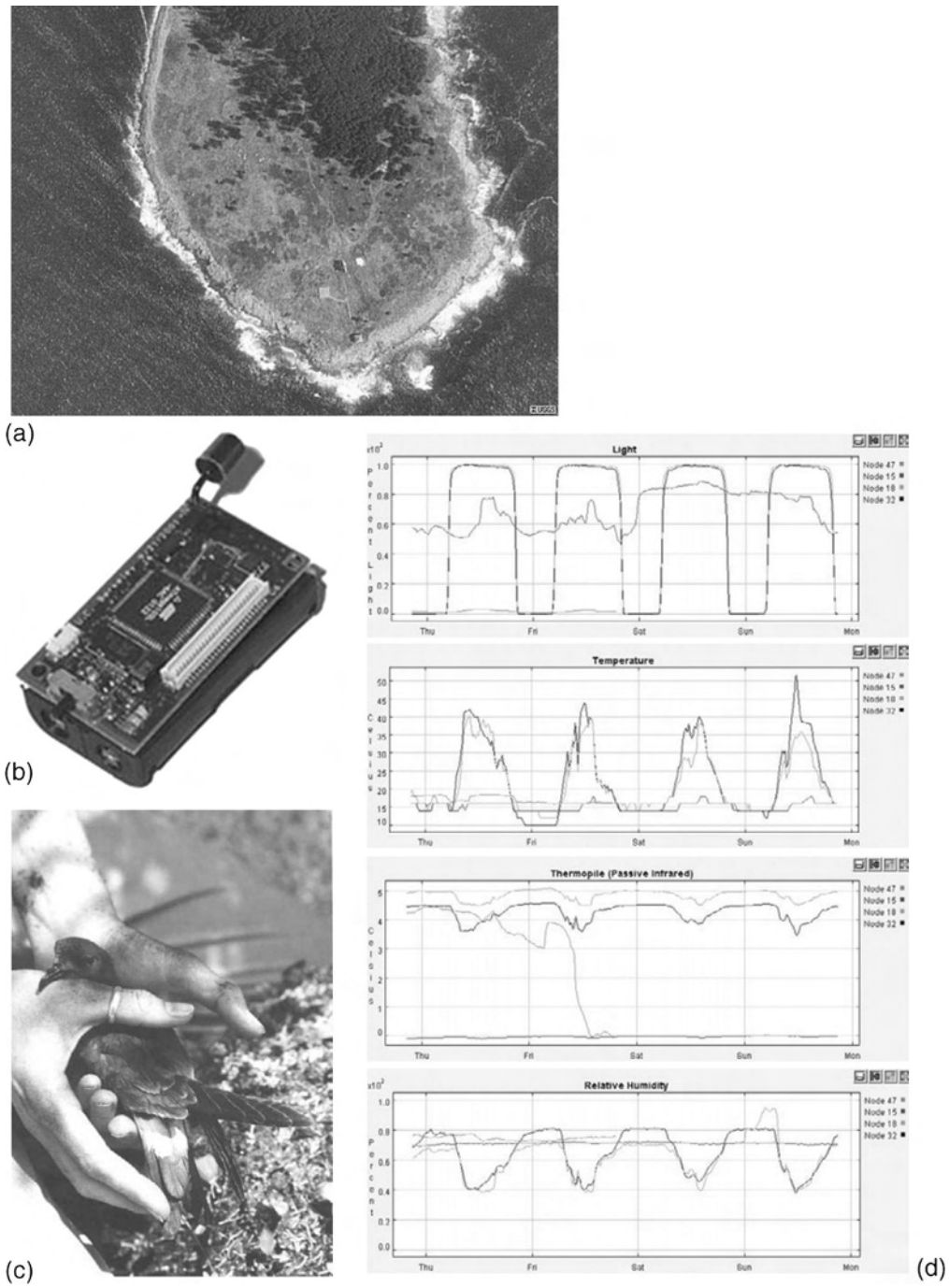


Figure 3.7-4 Remote wireless sensors are used to find out more about birds in their natural habitat. (a) Sensor network deployment on Great Duck Island, 10 miles off the coast of Maine, (b) Wireless biological sensors placed in bird nests, (c) A petrel, rarely seen by birdwatchers. (d) Sensor measurements remotely retrieved using a Web interface, summer 2002. (*Picture courtesy of U.S. Geological Survey, Intel Corporation, and John Anderson*).

petrels, they literally had to stick their hands into the burrows, often causing the birds to abandon their homes.

Now, these same biologists are checking on the birds on the island in the comfort of their offices, browsing data from the sensors linked by satellite. And their colleagues thousands of miles away can share the same experience, thanks to the Internet. The untethered, matchbox-sized sensors left in the burrows monitor the occupancy by recording temperature variations inside and wirelessly send the data to a gateway node on the island. Convenience aside, the more significant benefit of the technology is the minimization of disturbance to

the very habitat that scientists are trying to help preserve.

The experiment on Great Duck Island is a small lens into an expansive future. To grasp what might happen, multiply these 190 sensors by 10 million or 100 million and distribute them globally. When the sensor grid becomes ubiquitous, an enormous digital retina will stretch over the surface of the planet. This planet-scale system could help us understand and address tomorrow's environmental challenges, ranging from monitoring global biodiversity to sensing millions of low-level, nonpoint sources of pollution.

3.7.3.2 Tracking chemical plumes: Ad hoc, Just-in-Time deployment for mitigating disasters

Image the following scenario. The Valley Authority has just declared a region-wide emergency: A large-scale hazardous chemical gas leak occurred at a chemical processing plant twenty minutes ago. The National Guard has been activated to evacuate nearby towns and to close roads and bridges. To get a real-time situational assessment of the extent and movement of the gas release and help plan the evacuation, the SensorNet SWAT Team is called in. Three unmanned aerial vehicles (UAVs) are immediately launched from an open field 15 miles south of the accident site, each carrying 1000 tiny wireless chemical sensing nodes (see Figure 3.7-5). Upon flying over the vicinity of the accident site, the sensor nodes are air-dropped. While airborne, the nodes self-organize into an ad hoc network and relay the sensing result back to the UAVs nearby: Where is the plume? How big is it? What is the shape? How fast is it moving?

While the full realization of this scenario may take several years, pieces of the sensor network technology have already been demonstrated. For example, in March 2001, Kris Pister and his Berkeley team deployed wireless sensors using a UAV for a vehicle tracking field demonstration in the Mojave Desert of Southern California. Each sensor node is equipped with a microprocessor, an RF modem, and a magnetometer. Including the battery, the sensor node weighs about an ounce. The UAV air-dropped six wireless sensors along a road (Figure 3.7-6). Once on the ground, the sensors immediately self-organized into a network, initiated detection when a vehicle passed by, and relayed the tracking results back to the UAV.

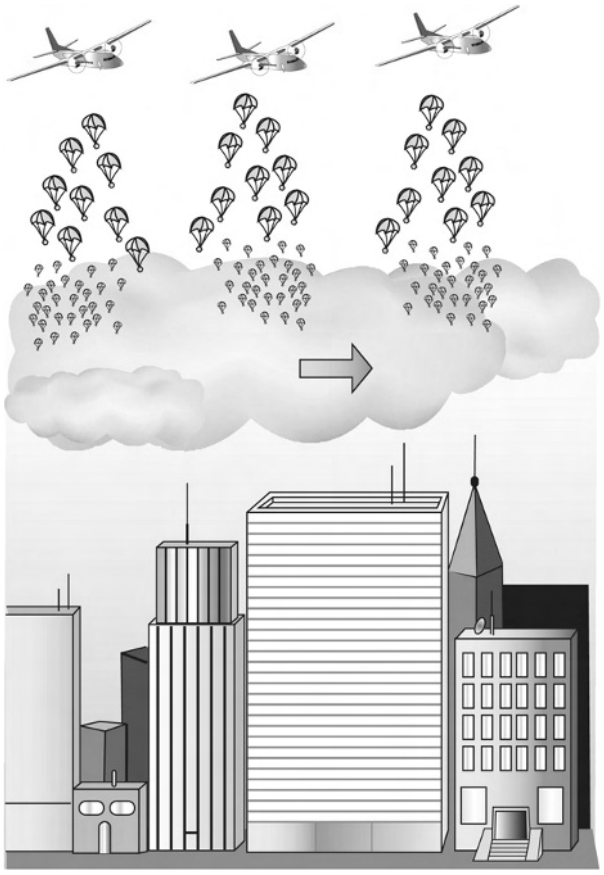


Figure 3.7-5 Tracking chemical plumes using ad hoc wireless sensors, deployed from air vehicles.

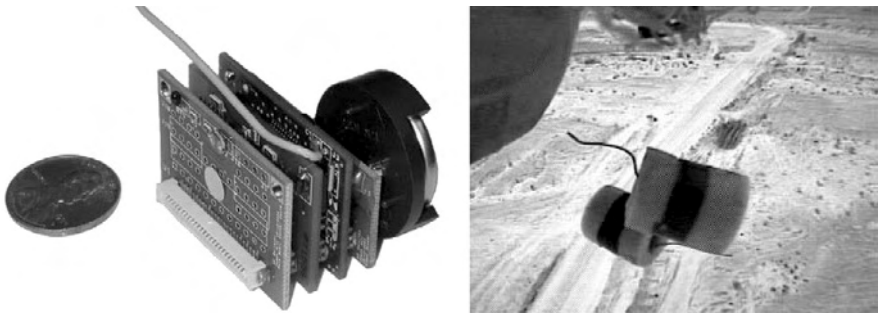


Figure 3.7-6 Left: Berkeley wireless sensor mote. Right: Air-drop of six sensor nodes from a UAV. (*Picture courtesy of Kris Pister and Jason Hill*).

3.7.3.3 Smart transportation: Networked Sensors making roads safer and less congested

Plenty of sensors are already in use for traffic monitoring purposes (see Figure 3.7-7). Sensors embedded in roadbeds or alongside highways measure traffic flow. Cameras at street intersections look for traffic violations. Sensors in vehicles monitor speed and other conditions. But today these sensors do not talk to each other as often as we would like them to.

When these sensors are networked together to share real-time information, we can begin to create a dynamic infrastructure for smart roads that can be optimized to make roads safer, reduce congestion, or help people find the nearest available parking space in an unfamiliar city. For example, cars and trucks equipped with wireless sensors can warn each other of imminent collision or other road hazards ahead. They can dynamically optimize routes to avoid traffic hot spots. They can continuously monitor their own conditions and transmit emission data to a service station to enable just-in-time maintenance.

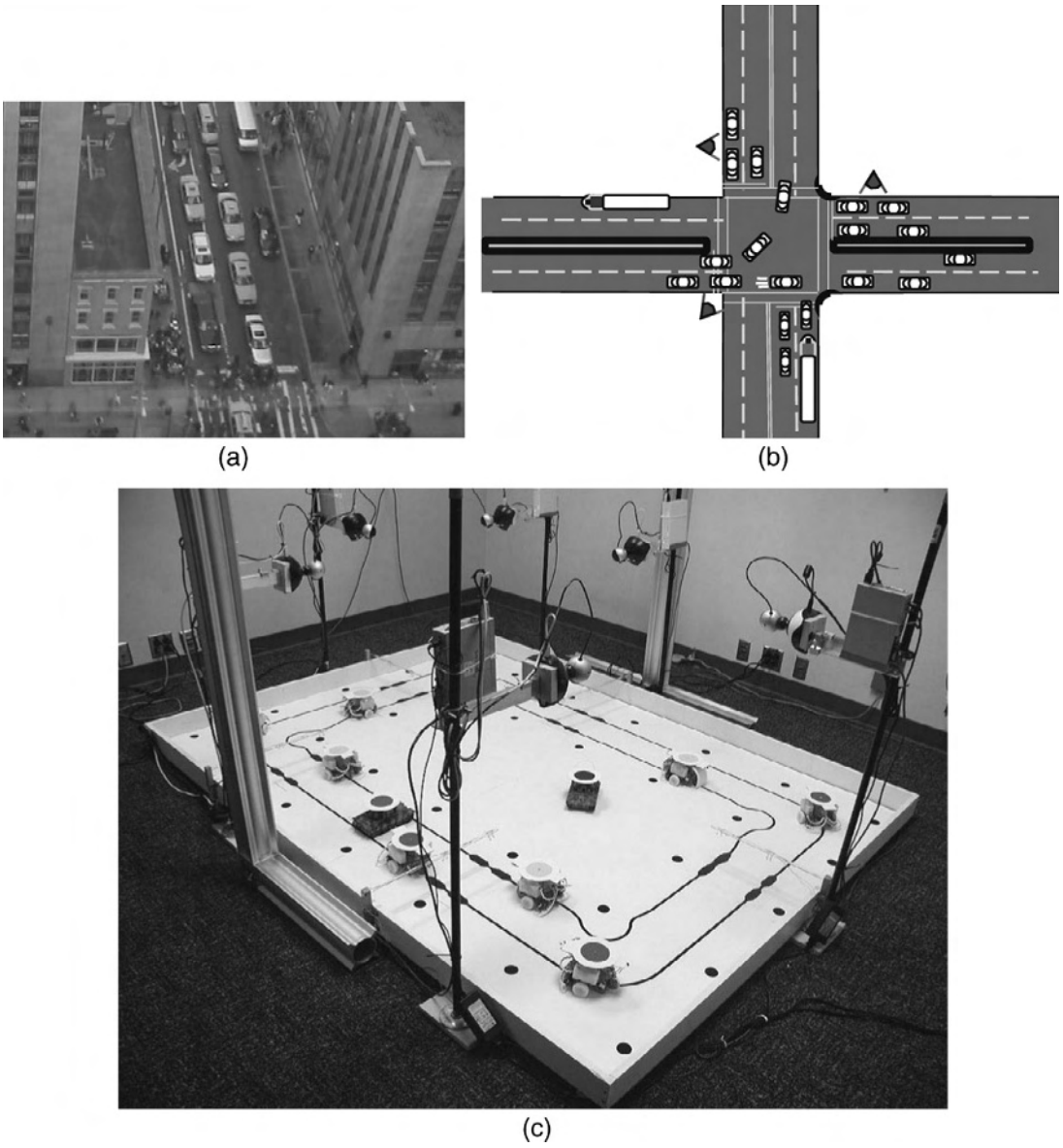


Figure 3.7-7 Distributed video sensor networks for traffic and security applications. Upper figures: Networked cameras and other sensors could be used to monitor traffic flow to reduce congestion, track vehicles on city streets for traffic violations, or detect illegal activities around critical infrastructure such as airports. Lower figure: PARC video sensor network prototype uses in-network intelligence to decide what events to pay attention to and what to ignore, thus reducing the amount of information the network must collect and transport in order to support high-level monitoring applications.

They can use the wireless connection to download music and movies while waiting at a gas station. For some of these applications, the crucial difference lies in the real-time flow of information. Point-to-point routing assistance or road safety application requires access to up-to-date or even instantaneous traffic data.

When sensors are used to guard critical infrastructure such as power plants or airports, they can provide a virtual enclosure around the facilities to guard against unauthorized intrusion. Unlike existing security systems that require human personnel to watch video feeds around the clock and pick out "unusual" events, which is clearly not scalable, smart sensor networks will have to rely on in-network intelligence to focus on interesting events, filter out detractors, extract meaning from raw sensor readings or video streams, and transmit only relevant scene features to human users.

3.7.4 Collaborative processing

As the above applications have illustrated, many sensing tasks require a sensor network system to process data cooperatively and to combine information from multiple sources. In traditional centralized sensing and signal processing systems, raw data collected by sensors are relayed to the edges of a network where the data is processed. From the scalability point of view, the nonlocal processing at the edges depletes precious bandwidth. If every sensor has some data that it needs to send to another node in a network, then a well-known wireless capacity result by Gupta and Kumar states that the per node throughput scales as $\frac{1}{\sqrt{N}}$—in other words, it goes to zero as the number of nodes N in a wireless ad hoc network increases. This result holds regardless of optimality in routing, power control, or transmission. Intuitively, this states that, as the number of nodes increases, every node spends almost all of its time forwarding packets of other nodes. From the energy point of view, transmitting raw data to distant nodes is wasteful of scarce resources. The diminishing wireless capacity result can be somewhat mitigated by introducing mobility to nodes, if an application is delay-tolerant.

In a sensor network context, one can clearly do better. Since data from multiple sensors with overlapping sensing regions is almost always correlated, one can remove the redundant information in the data, through in-network aggregation and compression local to the nodes that generate the data, before shipping it to a remote node. In fact, it can be shown that the amount of nonredundant data that a network generates grows as $O\left(\frac{\log N}{N}\right)$, assuming that the network is sampling a physical phenomenon with a prescribed accuracy requirement. This is encouraging since the amount of data generated per node scales as $O\left(\frac{\log N}{N}\right)$, which is within the per-node throughput constraint derived by Gupta and Kumar. On the other hand, instead of applying data compression techniques to samples after they are collected, nodes can be more selective in what data to generate or communicate. For energy-constrained and multiuser decentralized systems, it becomes critical to carefully select the sensor nodes that participate in a sensor collaboration, balancing the information contribution of each against its resource consumption or potential utility for other users. We use the term *collaborative signal and information processing* (CSIP) to refer to signal and information processing problems dominated by this issue of selecting embedded sensors to participate in an information processing task. The topic of active control and tasking of sensors is quite unique to sensor networks.

3.7.5 Key definitions of sensor networks

Sensor networks is an interdisciplinary research area that draws on contributions from signal processing, networking and protocols, databases and information management, distributed algorithms, and embedded systems and architecture. In the following, we define a number of key terms and concepts that will be used throughout the book as we develop techniques and examples for sensor networks.

- *Sensor:* A transducer that converts a physical phenomenon such as heat, light, sound, or motion into electrical or other signals that may be further manipulated by other apparatus.
- *Sensor node:* A basic unit in a sensor network, with on-board sensors, processor, memory, wireless modem, and power supply. It is often abbreviated as *node*. When a node has only a single sensor on board, the node is sometimes also referred to as a *sensor*, creating some confusion.
- *Network topology:* A connectivity graph where nodes are sensor nodes and edges are communication links. In a wireless network, the link represents a one-hop connection, and the neighbors of a node are those within the radio range of the node.
- *Routing:* The process of determining a network path from a packet source node to its destination.
- *Date-centric:* Approaches that name, route, or access a piece of data via properties, such as physical location, that are external to a communication network. This is to be contrasted with address-centric approaches which use logical properties of nodes related to the network structure.

- *Geographic routing:* Routing of data based on geographical attributes such as locations or regions. This is an example of date-centric networking.
- *In-network:* A style of processing in which the data is processed and combined near where the data is generated.
- *Collaborative processing:* Sensors cooperatively processing data from multiple sources in order to serve a high-level task. This typically requires communication among a set of nodes.
- *State:* A snapshot about a physical environment (e.g., the number of signal sources, their locations or spatial extent, speed of movement), or a snapshot of the system itself (e.g.,the network state).
- *Uncertainty:* A condition of the information caused by noise in sensor measurements, or lack of knowledge in models. The uncertainty affects the system's ability to estimate the state accurately and must be carefully modeled. Because of the ubiquity of uncertainty in the data, many sensor network estimation problems are cast in a statistical framework. For example, one may use a covariance matrix to characterize the uncertainty in a Gaussian-like process or more general probability distributions for non-Gaussian processes.
- *Task:* Either high-level system tasks which may include sensing, communication, processing, and resource allocation, or application tasks which may include detection, classification, localization, or tracking.
- *Detection:* The process of discovering the existence of a physical phenomenon. A threshold-based detector may flag a detection whenever the signature of a physical phenomenon is determined to be significant enough compared with the threshold.
- *Classification:* The assignment of class labels to a set of physical phenomena being observed.
- *Localization and tracking:* The estimation of the state of a physical entity such as a physical phenomenon or a sensor node from a set of measurements. Tracking produces a series of estimates over time.
- *Value of information or information utility:* A mapping of data to a scalar number, in the context of the overall system task and knowledge. For example, information utility of a piece of sensor data may be characterized by its relevance to an estimation task at hand and computed by a mutual information function.
- *Resource:* Resources include sensors, communication links, processors, on-board memory, and node energy reserves. Resource allocation assigns resources to tasks, typically optimizing some performance objective.
- *Sensor tasking:* The assignment of sensors to a particular task and the control of sensor state (e.g., on/off, pan/tilt) for accomplishing the task.
- *Node services:* Services such as time synchronization and node localization that enable applications to discover properties of a node and the nodes to organize themselves into a useful network.
- *Data storage:* Sensor information is stored, indexed, and accessed by applications. Storage may be local to the node where the data is generated, load-balanced across a network, or anchored at a few points (warehouses).
- *Embedded operating system (OS):* The run-time system support for sensor network applications. An embedded OS typically provides an abstraction of system resources and a set of utilities.
- *System performance goal:* The abstract characterization of system properties. Examples include scalability, robustness, and network longevity, each of which may be measured by a set of evaluation metrics.
- *Evaluation metric:* A measurable quantity that describes how well the system is performing on some absolute scale. Examples include packet loss (system), network dwell time (system), track loss (application), false alarm rate (application), probability of correct association (application), location error (application), or processing latency (application/system). An evaluation method is a process for comparing the value of applying the metrics on an experimental system with that of some other benchmark system.

Section **Four**

Mobile communications

This page intentionally left blank

Chapter 4.1

4.1

Background of 3G evolution

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

From the first experiments with radio communication by Guglielmo Marconi in the 1890s, the road to truly mobile radio communication has been quite long. To understand the complex third-generation (3G) mobile-communication systems of today, it is also important to understand where they came from and how cellular systems have evolved from an expensive technology for a few selected individuals to today's global mobile-communication systems used by almost half of the world's population. Developing mobile technologies has also changed, from being a national or regional concern, to becoming a very complex task undertaken by global standards-developing organizations such as the *Third Generation Partnership Project* (3GPP) and involving thousands of people.

4.1.1 History and background of 3G

The cellular technologies specified by 3GPP are the most widely deployed in the world, with the number of users passing 2 billion in 2006. The latest step being studied and developed in 3GPP is an evolution of 3G into an evolved radio access referred to as the *Long-Term Evolution* (LTE) and an evolved packet access core network in the *System Architecture Evolution* (SAE). By 2009–2010, LTE and SAE are expected to be deployed first.

Looking back to when it all it started, it begun several decades ago with early deployments of analog cellular services.

4.1.1.1 Before 3G

The US *Federal Communications Commission* (FCC) approved the first commercial car-borne telephony service in 1946, operated by AT&T. In 1947 AT&T also introduced the cellular concept of reusing radio frequencies, which became fundamental to all subsequent mobile-communication systems. Commercial mobile telephony continued to be car-borne for many years because of bulky and power-hungry equipment. In spite of the limitations of the service, there were systems deployed in many countries during the 1950s and 1960s, but the users counted only in thousands at the most.

These first steps on the road of mobile communication were taken by the monopoly telephone administrations and wire-line operators. The big uptake of subscribers and usage came when mobile communication became an international concern and the industry was invited into the process. The first international mobile communication system was the analog NMT system (*Nordic Mobile Telephony*) which was introduced in the Nordic countries in 1981, at the same time as analog AMPS (*Advanced Mobile Phone Service*) was introduced in North America. Other analog cellular technologies deployed worldwide were Total Access Communications System (TACS) and Japanese-Total Access Communication System (J-TACS). They all had in common that the equipment was still bulky, mainly car-borne, and voice quality was often inconsistent, with 'cross-talk' between users being a common problem.

With an international system such as NMT came the concept of 'roaming,' giving a service for users traveling outside the area of their 'home' operator. This also gave a larger market for the mobile phones, attracting more companies into the mobile communication business.

The analog cellular systems supported 'plain old telephony services,' that is voice with some related supplementary services. With the advent of digital communication during the 1980s, the opportunity to develop a second generation of mobile-communication standards and systems, based on digital technology, surfaced. With digital technology came an opportunity to

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved

increase the capacity of the systems, to give a more consistent quality of the service, and to develop much more attractive truly mobile devices.

In Europe, the telecommunication administrations in CEPT[1] initiated the GSM project to develop a pan-European mobile-telephony system. The GSM activities were in 1989 continued within the newly formed *European Telecommunication Standards Institute* (ETSI). After evaluations of TDMA, CDMA, and FDMA-based proposals in the mid-1980s, the final GSM standard was built on TDMA. Development of a digital cellular standard was simultaneously done by TIA in the USA resulting in the TDMA-based IS-54 standard, later simply referred to as US-TDMA. A somewhat later development of a CDMA standard called IS-95 was completed by TIA in 1993. In Japan, a second-generation (2G) TDMA standard was also developed, usually referred to as PDC.

All these standards were 'narrowband' in the sense that they targeted 'low-bandwidth' services such as voice. With the 2G digital mobile communications came also the opportunity to provide data services over the mobile-communication networks. The primary data services introduced in 2G were text messaging (SMS) and circuit-switched data services enabling e-mail and other data applications. The peak data rates in 2G were initially 9.6 kbps. Higher data rates were introduced later in evolved 2G systems by assigning multiple time slots to a user and by modified coding schemes.

Packet data over cellular systems became a reality during the second half of the 1990s, with *General Packet Radio Services* (GPRS) introduced in GSM and packet data also added to other cellular technologies such as the Japanese PDC standard. These technologies are often referred to as 2.5G. The success of the wireless data service iMode in Japan gave a very clear indication of the potential for applications over packet data in mobile systems, in spite of the fairly low data rates supported at the time.

With the advent of 3G and the higher-bandwidth radio interface of UTRA (*Universal Terrestrial Radio Access*) came possibilities for a range of new services that were only hinted at with 2G and 2.5G. The 3G radio access development is today handled in 3GPP. However, the initial steps for 3G were taken in the early 1990s, long before 3GPP was formed.

What also set the stage for 3G was the internationalization of cellular standardization. GSM was a pan-European project, but quickly attracted worldwide interest when the GSM standard was deployed in a number of countries outside Europe. There are today only three countries worldwide where GSM is not deployed. A global standard gains in economy of scale, since the market for products becomes larger. This has driven a much tighter international cooperation around 3G cellular technologies than for the earlier generations.

4.1.1.2 Early 3G discussions

Work on a 3G mobile communication started in ITU (*International Telecommunication Union*) in the 1980s. The radiocommunication sector ITU-R issued a first recommendation defining *Future Public Land Mobile Telecommunications Systems* (FPLMTS) in 1990, later revised in 1997. The name for 3G within ITU had by then changed from FPLMTS to IMT-2000. The World Administrative Radio Congress WARC-92 identified 230 MHz of spectrum for IMT-2000 on a worldwide basis. Of these 230 MHz, 2 × 60 MHz were identified as paired spectrum for FDD and 35 MHz as unpaired spectrum for TDD, both for terrestrial use. Some spectrum was also set aside for satellite services. With that, the stage was set to specify IMT-2000.

Task Group 8/1 within ITU-R developed a range of recommendations for IMT-2000, defining a framework for services, network architectures, radio interface requirements, spectrum considerations, and evaluation methodology. Both a terrestrial and a satellite component were defined.

Task Group 8/1 defined the process for evaluating IMT-2000 technologies in ITU-R recommendation M.1225. The evaluation criteria set the target data rates for the 3G circuit-switched and packet-switched data services:

- Up to 2 Mbps in an indoor environment.
- Up to 144 kbps in a pedestrian environment.
- Up to 64 kbps in a vehicular environment.

These numbers became the benchmark that all 3G technologies were compared with. However, already today, data rates well beyond 2 Mbps can be seen in deployed 3G systems.

4.1.1.3 Research on 3G

In parallel with the widespread deployment and evolution of 2G mobile-communication systems during the 1990s, substantial efforts were put into 3G research activities. In Europe the partially EU-funded project *Research into Advanced Communications in Europe* (RACE) carried out initial 3G research in its first phase. 3G in Europe was named *Universal Mobile Telecommunications Services* (UMTS). In the second phase of RACE, the CODIT project (*Code Division Test bed*) and the ATDMA project (*Advanced TDM A Mobile Access*) further developed 3G concepts based on *Wideband*

[1] *The European Conference of Postal and Telecommunications Administrations* (CEPT) consist of the telecom administrations from 47 countries.

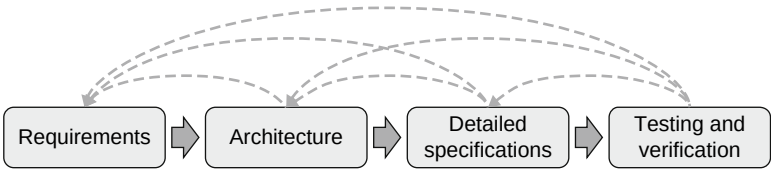


Figure 4.1-1 The standardization phases and iterative process.

CDMA (WCDMA) and Wideband TDMA technologies. The next phase of related European research was *Advanced Communication Technologies and Services* (ACTS), which included the UMTS-related project *Future Radio Wideband Multiple Access System* (FRAMES). The FRAMES project resulted in a multiple access concept that included both Wideband CDMA and Wideband TDMA components.

At the same time parallel 3G activities were going on in other parts of the world. In Japan, the *Association of Radio Industries and Businesses* (ARIB) was in the process of defining a 3G wireless communication technology based on Wideband CDMA. In the US, a Wideband CDMA concept called WIMS was developed within the T1.P1[2] committee. Korea also started work on Wideband CDMA at this time.

The FRAMES concept was submitted to the standardization activities for 3G in ETSI,[3] where other multiple access proposals were also introduced by the industry, including the Wideband CDMA concept from the ARIB standardization in Japan. The ETSI proposals were merged into five concept groups, which also meant that the Wideband CDMA proposals from both Europe and Japan were merged.

4.1.1.4 3G standardization starts

The outcome of the ETSI process in early 1998 was the selection of *Wideband CDMA* (WCDMA) as the technology for UMTS in the paired spectrum (FDD) and TD-CDMA (*Time Division CDMA*) for the unpaired spectrum (TDD). There was also a decision to harmonize the parameters between the FDD and the TDD components.

The standardization of WCDMA went on in parallel in ETSI and ARIB until the end of 1998 when the *Third Generation Partnership Project* (3GPP) was formed by standards-developing organizations from all regions of the world. This solved the problem of trying to maintain parallel development of aligned specifications in multiple regions. The present organizational partners of 3GPP are ARIB (Japan), CCSA (China), ETSI (Europe), ATIS (USA), TTA (Korea) and TTC (Japan).

4.1.2 Standardization

4.1.2.1 The standardization process

Setting a standard for mobile communication is not a one-time job, it is an ongoing process. The standardization forums are constantly evolving their standards trying to meet new demands for services and features. The standardization process is different in the different forums, but typically includes the four phases illustrated in Figure 4.1-1:

1. *Requirements*, where it is decided what is to be achieved by the standard.
2. *Architecture*, where the main building blocks and interfaces are decided.
3. *Detailed specifications*, where every interface is specified in detail.
4. *Testing and verification*, where the interface specifications are proven to work with real-life equipment.

These phases are overlapping and iterative. As an example, requirements can be added, changed, or dropped during the later phases if the technical solutions call for it. Likewise, the technical solution in the detailed specifications can change due to problems found in the testing and verification phase.

Standardization starts with the *requirements* phase, where the standards body decides what should be achieved with the standard. This phase is usually relatively short.

In the *architecture* phase, the standards body decides about the architecture, i.e. the principles of how to meet the requirements. The architecture phase includes decisions about reference points and interfaces to be standardized. This phase is usually quite long and may change the requirements.

After the architecture phase, the *detailed specification* phase starts. It is in this phase that the details for each of the identified interfaces are specified. During the detailed specification of the interfaces, the standards body may find that it has to change decisions made either in the architecture or even in the requirements phase.

[2] The T1.P1 committee was part of T1 which presently has joined the ATIS standardization organization.
[3] The TDMA part of the FRAMES project was also fed into 2G standardization as the evolution of GSM into EDGE (Enhanced Data rates for GSM Evolution).

Finally, the *testing and verification* phase starts. It is usually not a part of the actual standardization in the standards bodies, but takes place in parallel through testing by vendors and interoperability testing between vendors. This phase is the final proof of the standard. During the testing and verification phase, errors in the standard may still be found and those errors may change decisions in the detailed standard. Albeit not common, changes may need to be done also to the architecture or the requirements. To verify the standard, products are needed. Hence, the implementation of the products starts after (or during) the detailed specification phase. The testing and verification phase ends when there are stable test specifications that can be used to verify that the equipment is fulfilling the standard.

Normally, it takes one to two years from the time when the standard is completed until commercial products are out on the market. However, if the standard is built from scratch, it may take longer time since there are no stable components to build from.

4.1.2.2 3GPP

The *Third-Generation Partnership Project* (3GPP) is the standards-developing body that specifies the 3G UTRA and GSM systems. 3GPP is a partnership project formed by the standards bodies ETSI, ARIB, TTC, TTA, CCSA and ATIS. 3GPP consists of several *Technical Specifications Groups* (TSGs) (see Figure 4.1-2).

A parallel partnership project called 3GPP2 was formed in 1999. It also develops 3G specifications, but for cdma2000, which is the 3G technology developed from the 2G CDMA-based standard IS-95. It is also a global project, and the organizational partners are ARIB, CCSA, TIA, TTA and TTC.

3GPP TSG RAN is the technical specification group that has developed WCDMA, its evolution HSPA, as well as LTE, and is in the forefront of the technology. TSG RAN consists of five working groups (WGs):

1. RAN WG1 dealing with the physical layer specifications.
2. RAN WG2 dealing with the layer 2 and layer 3 radio interface specifications.
3. RAN WG3 dealing with the fixed RAN interfaces, for example interfaces between nodes in the RAN, but also the interface between the RAN and the core network.
4. RAN WG4 dealing with the *radio frequency* (RF) and *radio resource management* (RRM) performance requirements.
5. RAN WG5 dealing with the terminal conformance testing.

The scope of 3GPP when it was formed in 1998 was to produce global specifications for a 3G mobile system based on an evolved GSM core network, including the WCDMA-based radio access of the UTRA FDD and the

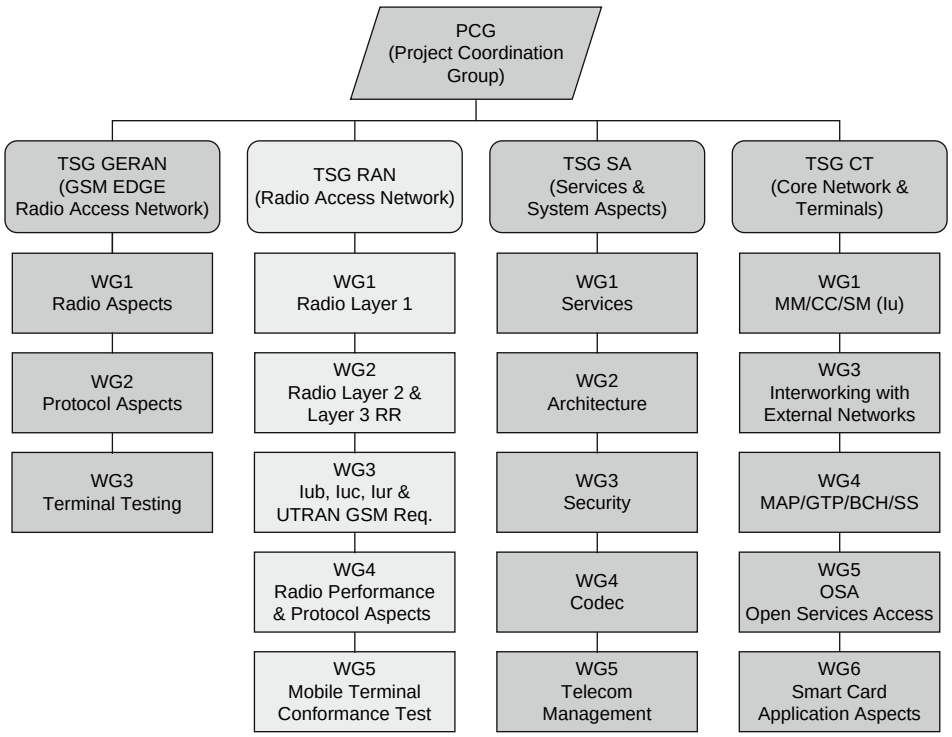


Figure 4.1-2 3GPP organization.

TD-CDMA-based radio access of the UTRA TDD mode. The task to maintain and develop the GSM/EDGE specifications was added to 3GPP at a later stage. The UTRA (and GSM/EDGE) specifications are developed, maintained and approved in 3GPP. After approval, the organizational partners transpose them into appropriate deliverables as standards in each region.

In parallel with the initial 3GPP work, a 3G system based on TD-SCDMA was developed in China. TD-SCDMA was eventually merged into Release 4 of the 3GPP specifications as an additional TDD mode.

The work in 3GPP is carried out with relevant ITU recommendations in mind and the result of the work is also submitted to ITU. The organizational partners are obliged to identify regional requirements that may lead to options in the standard. Examples are regional frequency bands and special protection requirements local to a region. The specifications are developed with global roaming and circulation of terminals in mind. This implies that many regional requirements in essence will be global requirements for all terminals, since a roaming terminal has to meet the strictest of all regional requirements. Regional options in the specifications are thus more common for base stations than for terminals.

The specifications of all releases can be updated after each set of TSG meetings, which occur 4 times a year. The 3GPP documents are divided into releases, where each release has a set of features added compared to the previous release. The features are defined in Work Items agreed and undertaken by the TSGs. The releases up to Release 7 and some main features of those are shown in Figure 4.1-3. The date shown for each release is the day the content of the release was frozen. For historical reasons, the first release is numbered by the year it was frozen (1999), while the following releases are numbered 4, 5, etc.

For the WCDMA Radio Access developed in TSG RAN, Release 99 contains all features needed to meet the IMT-2000 requirements as defined by ITU. There are circuit-switched voice and video services, and data services over both packet-switched and circuit-switched bearers. The first major addition of radio access features to WCDMA is Release 5 with *High Speed Downlink Packet Access* (HSDPA) and Release 6 with *Enhanced Uplink*. These two are together referred to as HSPA. With HSPA, UTRA goes beyond the definition of a 3G mobile system and also encompasses broadband mobile data.

With the studies of an Evolved UTRAN (LTE) and the related *System Architecture Evolution* (SAE), further steps are taken in terms of broadband capabilities.

4.1.2.3 IMT-2000 activities in ITU

The present ITU work on 3G takes place in ITU-R *Working Party 8F* (WP8F), where 3G systems are referred to as IMT-2000. WP8F does not write technical specifications for IMT-2000, but has kept the role of defining IMT-2000, cooperating with the regional standardization bodies and to maintain a set of recommendations for IMT-2000.

The main IMT-2000 recommendation is ITU-R M.1457, which identifies the IMT-2000 *radio interface specifications* (RSPC). The recommendation contains a 'family' of radio interfaces, all included on an equal basis. The family of five terrestrial radio interfaces is illustrated in Figure 4.1-4, which also shows what *Standards Developing Organizations* (SDO) or Partnership Projects produce the specifications. In addition, there are several IMT-2000 satellite radio interfaces defined, not illustrated in Figure 4.1-4.

For each radio interface, M.1457 contains an overview of the radio interface, followed by a list of references to the detailed specifications. The actual specifications are maintained by the individual SDOs and M.1457 provides

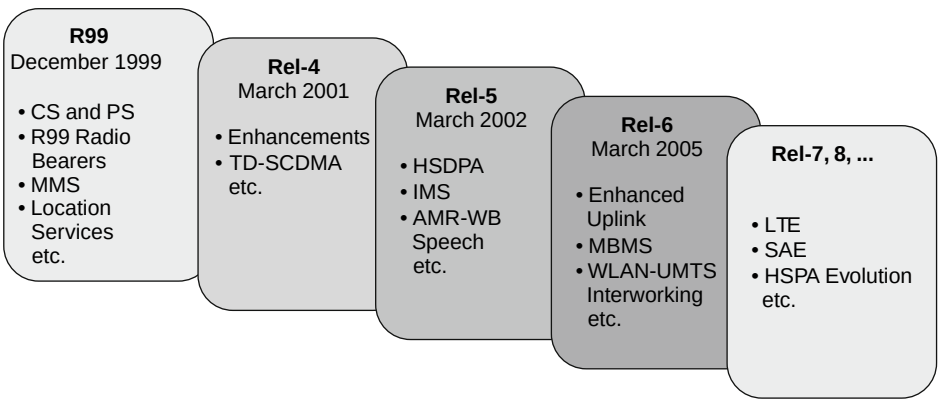


Figure 4.1-3 Releases of 3GPP specifications for UTRA.

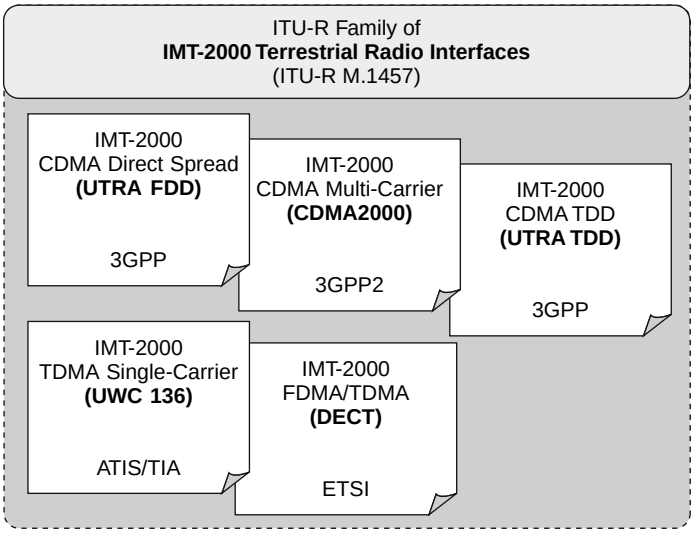


Figure 4.1-4 The definition of IMT-2000 in ITU-R.

URLs locating the specifications at each SDOs web archive.

With the continuing development of the IMT-2000 radio interfaces, including the evolution of UTRA to Evolved UTRA, the ITU recommendations also need to be updated. ITU-R WP8F continuously revises recommendation M.1457 and at the time of writing it is in its fifth version, with the sixth version awaiting approval. Input to the updates is provided by the SDOs and Partnership Projects writing the standards.

In addition to maintaining the IMT-2000 specifications, the main activity in ITU-R WP8F is the work on systems beyond IMT-2000, recently named IMT-Advanced. This work includes studies of services and technologies, market forecasts, principles for standardization, estimation of spectrum needs, and identification of candidate frequency bands for IMT-Advanced. The spectrum work also involves sharing studies between IMT-Advanced and other technologies in those bands.

4.1.3 Spectrum for 3G

Spectrum for 3G was first identified at the *World Administrative Radio Congress* WARC-92. Resolution 212 identified the bands 1885–2025 and 2110–2200 MHz as intended for use by national administrations that want to implement IMT-2000. Of these 230 MHz of 3G spectrum, 2 × 30 MHz were intended for the satellite component of IMT-2000 and the rest for the terrestrial component. Parts of the bands were (during the 1990s) used for deployment of 2G cellular systems, especially in the Americas. The first deployment of 3G in 2001–2002 by Japan and Europe were done in this band allocation, and it is for that reason often referred to as the IMT-2000 'core band.'

Additional spectrum for IMT-2000 was identified at the *World Radiocommunication Conference* WRC-2000 in Resolutions 223 and 224, where it was considered that an additional need for 160 MHz of spectrum for IMT-2000 was forecasted by ITU-R. The identification includes the bands used for 2G mobile systems in 806-960 MHz and 1710–1885 MHz, and 'new' 3G spectrum in the bands 2500–2690 MHz. The identification of bands assigned for 2G was also a recognition of the evolution of existing 2G mobile systems into 3G.

The somewhat diverging arrangement between regions of the frequency bands assigned to 3G means that there is not a single band that can be used for 3G roaming worldwide. Large efforts have however been put into defining a minimum set of bands that can be used to provide roaming. In this way, multi-band devices can provide efficient worldwide roaming for 3G.

The worldwide frequency arrangements are outlined in ITU-R recommendation M.1036. The recommendation also identifies which parts of the spectrum are paired and which are unpaired. For the paired spectrum, the bands for uplink (mobile transmit) and downlink (base-station transmit) are identified for *Frequency Division Duplex* (FDD) operation. The unpaired bands can for example be used for *Time Division Duplex* (TDD) operation. Note that the band that is most globally deployed for 3G is still 2 GHz.

3GPP first defined UTRA in Release 99 for the 2 GHz bands, with 2×60 MHz for UTRA FDD and 20+15 MHz of unpaired spectrum for UTRA TDD. A separate definition was also made for the use of UTRA in the US PCS bands at 1900 MHz. The concept of frequency bands with separate and release-independent

Table 4.1-1 Frequency bands defined by 3GPP for UTRA FDD.

FDD band	Uplink range (MHz)	Downlink range (MHz)	Main region(s)
I	1920–1980	2110–2170	Europe, Asia
II	1850–1910	1930–1990	Americas (Asia)
III	1710–1785	1805–1880	Europe, Asia (Americas)
IV	1710–1755	2110–2155	Americas
V	824–849	869–894	Americas
VI	830–840	875–885	Japan
VII	2500–2570	2620–2690	Europe
VIII	880–915	925–960	Europe, Asia
IX	1749.9–1784.9	1844.9–1879.9	Japan
X	1710–1770	2110–2170	Americas

Table 4.1-2 Frequency bands defined by 3GPP for UTRA TDD.

TDD band	Frequency range (MHz)	Main region(s) (MHz)
(a)	1900–1920 2010–2025	Europe, Asia
(b)	1850–1910 1930–1990	(Americas)
(c)	1910–1930	(Americas)
(d)	2570–2620	Europe

requirements were defined in Release 5 of the 3GPP specifications. The release-independence implies that a new frequency band added at a later release can be implemented also for earlier releases. All bands are also defined with consideration of what other bands may be deployed in the same region through special coexistence requirements for both base stations and terminals. These tailored requirements enable coexistence between 3G (and 2G) deployments in different bands in the same geographical area and even for co-location of base stations at the same sites using different bands.

Release 7 of the 3GPP specifications include ten frequency bands for FDD and four for TDD. The FDD bands are numbered with Roman numerals from I to X as shown in Table 4.1-1 and in Figure 4.1-5. The TDD bands have alphabetical designations (a)–(d) as shown in Table 4.1-2. With the release independence of the specifications, all bands can be implemented using only Release 99 functionality.

Some of the FDD bands are partly or fully overlapping. This is in most cases explained by regional differences in how the bands defined in M.1036 are implemented. At the same time, a high degree of commonality is desired to enable global roaming.

Band IV was introduced as a new band for the Americas following the addition of the 3G bands at WRC-2000. Its downlink overlaps completely with the

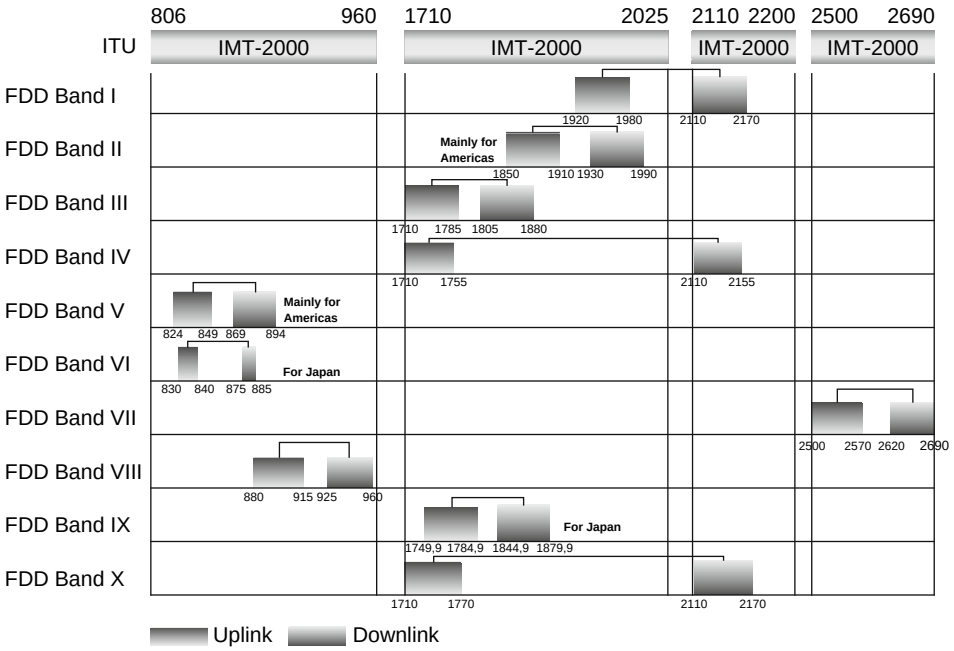


Figure 4.1-5 Operating bands for UTRA FDD specified in 3GPP.

downlink of Band I, which facilitates roaming and eases the design of dual band I/IV terminals.

Band X is an extension of Band IV from 2 × 45 to 2 × 60 MHz.

Bands V and VI overlap, but are intended for different regions, with Band VI restricted to Japan in the specifications. 2G systems in Japan had a very specific band plan and Band VI is the first step of aligning the Japanese spectrum plan in the 810–960 MHz range to that of other parts of the world.

Band IX overlaps with **Band III,** but is also intended only for Japan. The specifications are drafted in such a way that implementation of roaming dual band III/IX terminals is possible.

Chapter 4.2

4.2

The motives behind the 3G evolution

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

Before entering into a detailed discussion on technologies being used or considered for the evolution of 3G mobile communication, it is important to understand the motivation for this evolution: that is, understanding the underlying driving forces. This chapter will try to highlight some of the driving forces giving the reader an understanding of where the technical requirements and solutions are coming from.

4.2.1 Driving forces

A key factor for success in any business is to understand the forces that will drive the business in the future. This is in particular true for the mobile-communication industry, where the rapid growth of the number of subscribers and the global presence of the technologies have attracted several new players that want to be successful. Both new operators and new vendors try to compete with the existing operators and vendors by adopting new technologies and standards to provide new and existing services better and at a lower cost than earlier systems. The existing operators and vendors will, of course, also follow or drive new technologies to stay ahead of competition. Thus, there is a key driving force in staying competitive or becoming competitive.

From the technical perspective, the development in areas like digital cameras and color displays enables new fancier services than the existing mobile-communication services. To be able to provide those services, the mobile-communication systems need to be upgraded or even replaced by new mobile-communication technologies. Similarly, the technical advancement in digital processors enables new and more powerful systems that not only can provide the new services, but also can provide the existing successful services better and at a lower cost than the dominant mobile-communication technologies of today. Thus, the key drivers are:

- staying competitive;
- services (better provisioning of old services as well as provisioning of new services);
- cost (more cost-efficient provisioning of old services as well as cost-effective provisioning of new services).

The technology advancement is necessary to provide new and more advanced services at a reasonable cost as well as to provide existing services in a better and more cost-efficient way.

4.2.1.1 Technology advancements

Technology advancements in many areas make it possible to build devices that were not possible 20, 10, or even 5 years ago. Even though Moore's law[1] is not a law of physics, it gives an indication of the rapid technology evolution for integrated circuits. This evolution enables faster processing/computing in smaller devices at lower cost. Similarly, the rapid development of color screens, small digital cameras, etc. makes it possible to envisage services in a device that were seen as utopia 10 years ago. For an example of the terminal development in the past 20 years, see Figure 4.2-1.

[1] Moore's law is an empirical observation, and states that with the present rate of technological development, the complexity of an integrated circuit, with respect to minimum component cost, will double in about 18 months.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved



Figure 4.2-1 The terminal development has been rapid in the past 20 years.

The size and weight of the mobile terminals have been reduced dramatically during the past 20 years. The standby and talk times have also been extended dramatically and the end users do not need to re-charge their devices every day. Simple black-and-white (or brown-and-gray) numerical screens have evolved into color screens capable of showing digital photos at good quality. Mega-pixel-capable digital cameras have been added making the device more attractive to use. Thus, the mobile device has become a multi-purpose device, not only a mobile phone for voice communications.

In parallel to the technical development of the mobile devices, the mobile-communication technologies are developed to meet the demands of the new services enabled, and also to enable them wireless. The development of the digital signal processors enables more advanced receivers capable of processing megabits of data in a short time, and the introduction of the optical fibers enables high-speed network connections to the base stations. In sum, this enables a fast access to information on the Internet as well as a short roundtrip time for normal communications. Thus, new and fancier services are enabled by the technical development of the devices, and new and more efficient mobile-communication systems are enabled by a similar technical development.

4.2.1.2 Services

Delivering services to the end users is the fundamental goal of any mobile-communication system. Knowing them, understanding them, managing them, and charging them properly is the key for success. It is also the most difficult task being faced by the engineers developing the mobile-communication system of the future. It is very difficult to predict what service(s) will be popular in a 5- to 10-year perspective. In fact, the engineers have to design a system that can adapt to any service that might become popular and used in the future. Unfortunately, there are also technical limitations that need to be understood, and also the technical innovations that in the future enable new services.

4.2.1.2.1 Internet and IP technology

The Internet and the IP-based services delivered over the Internet are more and more going wireless. This means that the mobile-communication systems are delivering more and more IP-based services, from the best effort-Internet data to voice-over-IP, for example in the shape of push-to-talk (PoC). Furthermore, in the wireless environment it is more natural to use, for example, location-based services and tracking services than in the fixed environment. Thus, one can talk about mobile Internet services in addition to the traditional Internet services like browsing, e-mail, etc.[2]

Essentially, IP provides a transport mechanism that is service agnostic. Albeit there are several protocols on top of IP that are service-type specific (RTP, TCP, etc.), IP in itself is service agnostic. That enables service developers to develop services that only the imagination (and technology) sets the limit to. Thus, services will pop up, some will become popular for a while and then just fade away, whereas some others will never become a hit. Some services will become classics that will live and be used for a very long time.

4.2.1.2.2 Traditional telephony services

Going toward IP-based services obviously does not mean that traditional services that have been provided

[2] The common denominator between mobile Internet services and Internet services is the IP addressing technology with the IPv4 and IPv6 addresses identifying the end receiver. However, there is a need to handle the mobility provided by the cellular systems. Mobile IP is one possibility, but most of the cellular systems (if not all) have their own more efficient mobility mechanism.

over the circuit-switched domain, in successful mobile-communication systems like GSM, will disappear. Rather, it means that the traditional circuit-switched services will be ported over the IP networks. One particular service is the circuit-switched telephony service that will be provided as VoIP service instead. Thus, both the new advanced services that are enabled by the technology advancement of the devices and the traditional circuit-switched services will use IP as the transporting mechanism (and are therefore called IP-based services). Hence future mobile-communication networks, including the 3G evolution, need to be optimized for IP-based applications.

4.2.1.2.3 Wide spectrum of service needs

Trying to predict all the services that will be used over the mobile-communication systems 10 years from now is very difficult. The technology advancements in the various areas enable higher data rate connections, more memory on local devices, and more intelligent and easy to use man-machine interfaces. Furthermore, the human need of interaction and competition with other humans drives more intensive communication needs. All these combinations point toward applications and services that consume higher data rates and require lower delays compared to what today's mobile-communication systems can deliver.

However, the relatively low-rate voice service will still be a very important component of the service portfolio that mobile-network operators wish to provide. In addition, services that have very relaxed delay requirements will also be there. Thus, not only high data rate services with a low-latency requirement, but also low data rate best effort services will be provided. Furthermore, not only the data rate and delay are important to understand when talking about a service's need from a mobile-communication system, but also the setup time is very important. For example, a service can be totally useless if it takes too long to start it (for example, making a phone call, downloading a web page). Thus the mobile-communication systems of the future, including the evolution of 3G mobile communication, need to be able to deliver short call setup times, low latency and a wide range of data rates.

4.2.1.2.4 Key design services

Since it is impossible to know what services will be popular and since service possibilities and offers will differ with time and possibly also with country, the future mobile-communication systems will need to be adaptive to the changing service environment. Luckily, there are a few known key services that span the technology space. Those are:

- *Real-time-gaming applications:* These have the characteristic to require small amount of data (game update information) relatively frequent with low delay requirement.[3] Only a limited delay jitter is tolerable. A first person shooter game like Counter Strike is an example of a game that has this characteristic.
- *Voice:* This has the characteristic to require small amounts of data (voice packets) frequently with no delay jitter. The end-to-end delay has to be small enough not to be noticeable.[4]
- *Interactive file download and upload applications:* These have the characteristics of requiring low delay and high data rates.
- *Background file download and upload applications:* These have the characteristics of accepting lower bit rates and longer delays. E-mail (mostly) is an example of background file download and upload.
- *Television:* This has the characteristics of streaming downlink to many users at the same time requiring low delay jitter. The service can tolerate delays, as long as it is approximately the same delay for all users in the neighborhood. The television service has moderate data rate requirements.

A mobile-communication system designed to handle these services and the services in between will be able to facilitate most services (see Figure 4.2-2). Unfortunately, the upper limit of the data rate demand and the lower limit of the delay requirement are difficult to provide in a cost-efficient manner. The designers of the mobile-communication systems need to stop at a reasonable level, a level that the technology available at the time of standardization can provide.

4.2.1.3 Cost and performance

There is another important driving factor for future mobile-communication systems and that is the cost of the service provisioning. The technology advancement that enables new services can also be utilized to provide better mobile-communication systems using more advanced technical features. Here IP technology is not only a key enabler to allow for new services to pop up, but also a way of reducing cost of new services. The reason is that IP as a bearer of new services can be used to introduce new services as they come, not requiring an extensive

[3] The faster the data are delivered the better. Expert Counter Strike players look for game servers with a ping time of less than 50 ms.
[4] In 3G systems the end-to-end delay requirement for circuit-switched voice is approximately 400 ms. This delay is not disturbing humans in voice communications.

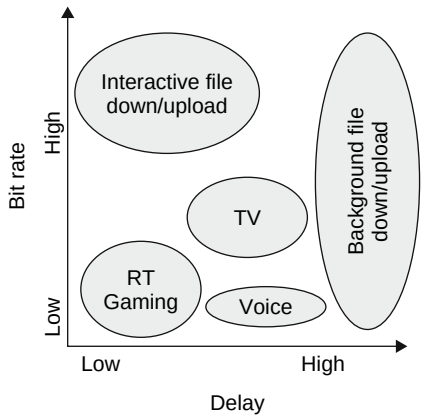


Figure 4.2-2 The bit rate – delay service space that is important to cover when designing a new cellular system.

special design of the system. Of course, this requires that the devices used in the mobile-communication system can be programmed by third-party providers and that the operators allow third-party service providers to use their network for communication.

Another important factor is that operators need to provide the services to all the users. Not only one user needs to get the low delay, high data rate, etc. that its service needs, but all the users with their different service needs should be served efficiently. The processing capacity evolution and Moore's law also help in this problem. New techniques are enabled by the higher processing power in the devices – techniques that deliver more bits of data per hertz. Furthermore, the coverage is increased with more advanced antennas and receivers. This enables the operators to deliver the services to more users from one base station, thus requiring fewer sites. Fewer sites imply lower operational and capitalization costs. In essence, the operators need fewer base stations and sites to provide the service.

Obviously, all services would be 'happy' if they were provided with the highest data rate, lowest delay, and lowest jitter that the system can provide. Unfortunately, this is unattainable in practice and contradictory to the operator goal of an efficient system: in other words, the more delay a service can handle the more efficient the system can be. Thus, the cost of providing lowest possible delay, jitter and call setup time is somewhat in conflict with the need of the mobile-network operator to provide it to all the users. Hence, there is a trade-off between user experience and system performance. The better the system performance is, the lower the cost of the network. However, the end users also need to get adequate performance which often is in conflict with the system performance, thus the operator cannot optimize for only system performance.

4.2.2 3G Evolution: Two radio access network approaches and an evolved core network

4.2.2.1 Radio access network evolution

TSG RAN organized a workshop on 3GPP long-term Evolution in the fall of 2004. The workshop was the starting point of the development of the Long-Term Evolution (LTE) radio interface. After the initial requirement phase in the spring of 2005, where the targets and objectives of LTE were settled, the technical specification group TSG SA launched a corresponding work on the System Architecture Evolution, since it was felt that the LTE radio interface needed a suitable evolved system architecture.

The result of the LTE workshop was that a study item in 3GPP TSG RAN was created in December 2004. The first 6 months were spent on defining the requirements, or design targets, for LTE. These were documented in a 3GPP technical report and approved in June 2005. Most notable are the requirements on high data rate at the cell edge and the importance of low delay, in addition to the normal capacity and peak data rate requirements. Furthermore, spectrum flexibility and maximum commonality between FDD and TDD solutions are pronounced.

During the fall 2005, 3GPP TSG RAN WG1 made extensive studies of different basic physical layer technologies and in December the TSG RAN plenary decided that the LTE radio access should be based on *OFDM* in the downlink and *single carrier FDMA* in the uplink.

At the same time as the LTE discussion was ongoing, 3GPP TSG RAN and its WGs continued to evolve the WCDMA system with more functionality, most notably MBMS and Enhanced Uplink. These additions were done in a backward compatible manner: that is, terminals of earlier releases can coexist on the same carrier in the same base station as terminals of the latest release. The main argument for the backward compatibility is that the installed base of equipment can be upgraded to handle the new features while still being capable of serving the old terminals. This is a cost-efficient addition of new features, albeit the new features are restricted by the solutions for the old terminals.

Naturally, HSPA[5] does not include all the technologies considered for LTE. Therefore, a study in 3GPP was initiated to see how far it is possible to take HSPA within the current spectrum allocation of 5 MHz and still

[5] When operating with HSDPA and Enhanced UL, the system is known as HSPA.

respect the backward compatibility aspect. Essentially, the target with HSPA Evolution is to reach near the characteristics of LTE when using a 5 MHz spectrum and at the same time being backward compatible.

Thus, 3GPP is working on two approaches for 3G evolution: the LTE and the HSPA Evolution. Both approaches have their merits. LTE can operate in new and more complex spectrum arrangements (although in the same spectrum bands as WCDMA and other 3G technologies) with the possibility for new designs that do not need to cater for terminals of earlier releases. HSPA Evolution can leverage on the installed base of equipment in the 5 MHz spectrum but needs to respect the backward compatibility of earlier terminals.

4.2.2.1.1 LTE drivers and philosophy

The 3GPP LTE is intended to be a mobile-communication system that can take the telecom industry into the 2020s. The philosophy behind LTE standardization is that the competence of 3GPP in specifying mobile-communication systems in general and radio interfaces in particular shall be used, but the result shall not be restricted by previous work in 3GPP. Thus, LTE do not need to be backward compatible with WCDMA and HSPA.

Leaving the legacy terminals behind, not being restricted by designs of the late 1990s, makes it possible to design a radio interface from scratch. In the LTE case, the radio interface can be purely optimized for IP transmissions not having to support ISDN traffic: that is, there is no requirement for support of GSM circuit-switch services, a requirement that WCDMA had. Furthermore, LTE also has the opportunity to maximize the commonality of FDD and TDD operations, a situation that did not exist in 3GPP before LTE.

Instead new requirements have arisen, for example the requirement on spectrum flexibility, since the global spectrum situation becomes more and more complex. Operators get more and more scattered spectrum, spread over different bands with different contiguous bandwidths. LTE needs to be able to operate in all these bands and with the bandwidths that is available to the operator. However, in practice only a limited set of bandwidths can be used since otherwise the RF and filter design would be too costly. LTE is therefore targeted to operate in 1.25, 1.6, 2.5, 5, 10, 15, and 20 MHz spectrum allocations. The spectrum flexibility support with the possibility to operate in other bandwidths than 5 MHz makes LTE very attractive for operators. The low-bandwidth operations are suitable for refarming of spectrum (for example, GSM spectrum and CDMA2000 spectrum). The higher-bandwidth options are suitable for new deployments in unused spectrum, where it is more common to have larger chunks of contiguous spectrum.

Furthermore, when going to the data rates that LTE is targeting, achieving low delay and high data rates at the cell edges are more important requirements than the peak data rate. Thus, a more pronounced requirement for LTE is the low delay with high data rates at the cell edges than it was when WCDMA was designed in the late 1990s.

Although not backward compatible with WCDMA, LTE design is clearly influenced by the WCDMA and the HSPA work in 3GPP. It is the same body, the same people, and companies that are active and more importantly, WCDMA and HSPA protocols are a good foundation for the LTE design. The philosophy is to take what is good from WCDMA and HSPA, and redo those parts that have to be updated due to the new requirement situation: either there are new requirements such as the spectrum flexibility or there are requirements that no longer are valid such as the support of ISDN services. The technology advancement in the cellular area has, of course, also influenced the design choice of LTE.

4.2.2.1.2 HSPA evolution drivers and philosophy

WCDMA, HSDPA, and soon also HSPA are in commercial operation throughout the world. This means that the infrastructure for HSPA is already in place with the network node sites, especially the base-station sites with their antenna arrangements and hardware. This equipment is serving millions of terminals with different characteristics and supported 3GPP releases. These terminals need to be supported by the WCDMA operators for many more years.

The philosophy of the HSPA Evolution work is to continue to add new and fancier technical features, and at the same time be able to serve the already existing terminals. This is the successful strategy of GSM that have added new features constantly since the introduction in the early 1990s. The success stems from the fact that there are millions of existing terminals at the launch time of the new features that can take the cost of the upgrade of the network for the initially few new terminals before the terminal fleet is upgraded. The time it takes to upgrade the terminal fleet is different from country to country, but a rule of thumb is that a terminal is used for 2 years before it is replaced. For HSPA Evolution that means that millions of HSPA-capable terminals need to be supported at launch. In other words, HSPA Evolution needs to be backward compatible with the previous releases in the sense that it is possible to serve terminals of earlier releases of WCDMA on the same carrier as HSPA-Evolution-capable terminals.

The backward compatibility requirement on the HSPA Evolution puts certain constraints on the technology that LTE does not need to consider, for example



Figure 4.2-3 One HSPA and LTE deployment strategy: upgrade to HSPA Evolution, then deploy LTE as islands in the WCDMA/HSPA sea.

the physical layer fundamentals need to be the same as for WCDMA release 99. In particular, the bandwidth needs to be the same (that is 5 MHz) which limits the deployment possibilities and puts a lower bound on the peak rate than LTE have the potential to deliver. On the other hand, HSPA Evolution is built on the existing specifications and only those parts of the specifications that need to be upgraded are touched. Thus there is less standardization, implementation and testing work for HSPA Evolution than LTE since the HSPA Evolution philosophy is to apply new more advanced technology on the existing HSPA standard. This will bring HSPA to a performance level comparable to LTE when compared on a 5 MHz spectrum allocation.

4.2.2.2 An evolved core network: System architecture evolution

Roughly at the same time as LTE and HSPA Evolution was started, 3GPP decided to make sure that an operator can coexist easily between HSPA Evolution and LTE through an evolved core network, the *Evolved Packet Core*. This work was done under the umbrella *System Architecture Evolution* study item lead by TSG SA WG2.

The *System Architecture Evolution* study focused on how the 3GPP core network will evolve into the core network of the next decades. The existing core network was designed in the 1980s for GSM, extended during the 1990s' for GPRS and WCDMA. The philosophy of the SAE is to focus on the packet-switched domain, and migrate away from the circuit-switched domain. This is done through the coming 3GPP releases ending up with the Evolved Packet Core.

Knowing that HSPA Evolution is backward compatible and knowing that the Evolved Packet Core will support both HSPA Evolution and LTE assures that LTE can be deployed in smaller islands and thus only where it is needed. A gradual deployment approach can be selected (see Figure 4.2-3). First the operator can upgrade its HSPA network to HSPA-Evolution-capable network, and then add LTE cells where capacity is lacking or where the operator wants to try out new services that cannot be delivered by HSPA Evolution. This approach reduces the cost of deployment since LTE do not need to be build for nationwide coverage from day one.

Chapter 4.3

4.3

High data rates in mobile communications

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

As discussed in Chapter 4.2, one main target for the evolution of 3G mobile communication is to provide the possibility for significantly higher end-user data rates compared to what is achievable with, for example, the first releases of the 3G standards. This includes the possibility for higher peak data rates but, as pointed out in the previous chapter, even more so the possibility for significantly higher data rates over the entire cell area, also including, for example, users at the cell edge. The initial part of this chapter will briefly discuss some of the more fundamental constraints that exist in terms of what data rates can actually be achieved in different scenarios. This will provide a background to subsequent discussions in the later part of the chapter concerning different means to increase the achievable data rates in different mobile-communication scenarios.

4.3.1 High data rates: Fundamental constraints

Shannon provided the basic theoretical tools needed to determine the maximum rate, also known as the *channel capacity*, by which information can be transferred over a given communication channel. Although relatively complicated in the general case, for the special case of communication over a channel, e.g. a radio link, only impaired by additative white Gaussian noise, the channel capacity C is given by the relatively simple expression

$$C = BW \cdot \log_2\left(1 + \frac{S}{N}\right) \tag{4.3.1}$$

where BW is the bandwidth available for the communication, S denotes the received signal power, and N denotes the power of the white noise impairing the received signal.

Already from (4.3.1) it should be clear that the two fundamental factors limiting the achievable data rate are the available received signal power, or more generally the available signal-power-to-noise-power ratio S/N, and the available bandwidth. To further clarify how and when these factors limit the achievable data rate, assume communication with a certain information rate R. The received signal power can then be expressed as $S = E_b \cdot R$ where E_b is the received energy per information bit. Furthermore, the noise power can be expressed as $N = N_0 \cdot \mathrm{BW}$ where N_0 is the constant noise power spectral density measured in W/Hz.

Clearly, the information rate can never exceed the channel capacity. Together with the above expressions for the received signal power and noise power, this leads to the inequality

$$\begin{aligned} R \le C &= BW \cdot \log_2\left(1 + \frac{S}{N}\right) \\ &= BW \cdot \log_2\left(1 + \frac{E_b \cdot R}{N_0 \cdot BW}\right) \end{aligned} \tag{4.3.2}$$

or, by defining the radio-link *bandwidth utilization* $\gamma = R/BW$,

$$\gamma \le \log_2\left(1 + \gamma \cdot \frac{E_b}{N_0}\right) \tag{4.3.3}$$

This inequality can be reformulated to provide a lower bound on the required received energy per information bit, normalized to the noise power density, for a given bandwidth utilization γ

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved

$$\frac{E_b}{N_0} \geq \min\left\{\frac{E_b}{N_0}\right\} = \frac{2^\gamma - 1}{\gamma} \qquad (4.3.4)$$

The rightmost expression, i.e. the minimum required E_b/N_0 at the receiver as a function of the bandwidth utilization is illustrated in Figure 4.3-1. As can be seen, for bandwidth utilizations significantly less than one, that is for information rates substantially smaller than the utilized bandwidth, the minimum required E_b/N_0 is relatively constant, regardless of γ. For a given noise power density, any increase of the information data rate then implies a similar relative increase in the minimum required signal power $S = E_b \cdot R$ at the receiver. On the other hand, for bandwidth utilizations larger than one the minimum required, E_b/N_0 increases rapidly with γ. Thus, in case of data rates in the same order as or larger than the communication bandwidth, any further increase of the information data rate, without a corresponding increase in the available bandwidth, implies a larger, eventually much larger, relative increase in the minimum required received signal power.

4.3.1.1 High data rates in noise-limited scenarios

From the discussion above, some basic conclusions can be drawn regarding the provisioning of higher data rates in a mobile-communication system when noise is the main source of radio-link impairment (a *noise-limited* scenario):

- The data rates that can be provided in such scenarios are always limited by the available received signal power or, in the general case, the received signal power-to-noise-power ratio. Furthermore, any increase of the achievable data rate within a given bandwidth will require *at least* the same relative increase of the received signal power. At the same time, if sufficient received signal power can be made available, basically any data rate can, at least in theory, be provided within a given limited bandwidth.
- In case of low-bandwidth utilization, i.e. as long as the radio-link data rate is substantially lower than the available bandwidth, any further increase of the data rate requires *approximately the same* relative increase in the received signal power. This can be referred to as *power-limited* operation (in contrast to *bandwidth-limited* operation, see below) as, in this case, an increase in the available bandwidth does not substantially impact what received signal power is required for a certain data rate.
- On the other hand, in case of high-bandwidth utilization, i.e. in case of data rates in the same order as or exceeding the available bandwidth, any further increase in the data rate requires *a much larger* relative increase in the received signal power unless the bandwidth is increased in proportion to the increase in data rate. This can be referred to *bandwidth-limited operation* as, in this case, an increase in the bandwidth will reduce the received signal power required for a certain data rate.

Thus, to make efficient use of the available received signal power or, in the general case, the available signal-to-noise ratio, the transmission bandwidth should at least be of the same order as the data rates to be provided.

Assuming a constant transmit power, the received signal power can always be increased by reducing the distance between the transmitter and the receiver, thereby reducing the attenuation of the signal as it propagates from the transmitter to the receiver. Thus, in a noise-limited scenario it is at least in theory always possible to increase the achievable data rates, assuming that one is prepared to accept a reduction in the

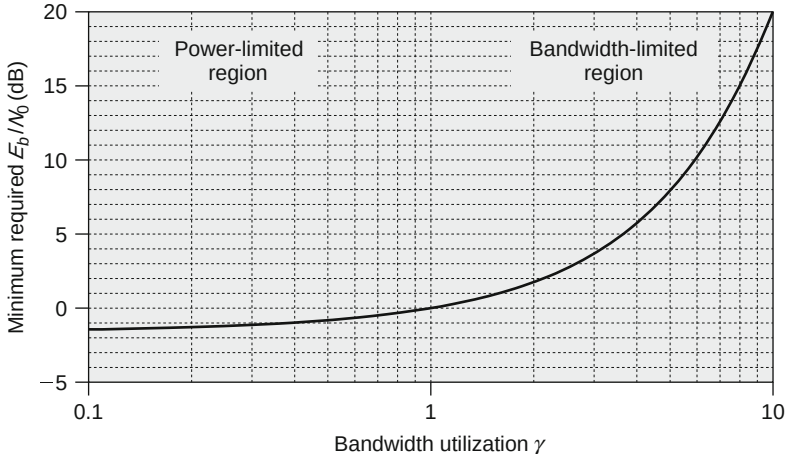


Figure 4.3-1 Minimum required E_b/N_0 at the receiver as a function of bandwidth utilization.

transmitter/receiver distance, that is a reduced range. In a mobile-communication system this would correspond to a reduced cell size and thus the need for more cell sites to cover the same overall area. Especially, providing data rates in the same order as or larger than the available bandwidth, i.e. with a high-bandwidth utilization, would require a significant cell-size reduction. Alternatively, one has to accept that the high data rates are only available for mobile terminals in the center of the cell, i.e. not over the entire cell area.

Another means to increase the overall received signal power for a given transmit power is the use of additional antennas at the receiver side, also known as *receive-antenna diversity*. Multiple receive antennas can be applied at the base station (that is for the uplink) or at the mobile terminal (that is for the downlink). By proper combining of the signals received at the different antennas, the signal-to-noise ratio after the antenna combining can be increased in proportion to the number of receive antennas, thereby allowing for higher data rates for a given transmitter/receiver distance.

Multiple antennas can also be applied at the transmitter side, typically at the base station, and be used to focus a given total transmit power in the direction of the receiver, i.e. toward the target mobile terminal. This will increase the received signal power and thus, once again, allow for higher data rates for a given transmitter/receiver distance.

However, providing higher data rates by the use of multiple transmit or receive antennas is only efficient up to a certain level, i.e. as long as the data rates are power limited rather than bandwidth limited. Beyond this point, the achievable data rates start to saturate and any further increase in the number of transmit or receive antennas, although leading to a correspondingly improved signal-to-noise ratio at the receiver, will only provide a marginal increase in the achievable data rates. This saturation in achievable data rates can be avoided though, by the use of multiple antennas at both the transmitter *and* the receiver, enabling what can be referred to as *spatial multiplexing*, often also referred to as *MIMO* (Multiple-Input Multiple-Output).

An alternative to increasing the received signal power is to reduce the noise power, or more exactly the noise power density, at the receiver. This can, at least to some extent, be achieved by more advanced receiver RF design, allowing for a reduced receiver noise figure.

4.3.1.2 Higher data rates in interference-limited scenarios

The discussion above assumed communication over a radio link only impaired by noise. However, in actual mobile-communication scenarios, interference from transmissions in neighbor cells, also referred to as *inter-cell interference*, is often the dominating source of radio-link impairment, more so than noise. This is especially the case in small-cell deployments with a high traffic load. Furthermore, in addition to inter-cell interference there may in some cases also be interference from other transmissions *within the current cell*, also referred to as *intra-cell interference*.

In many respects the impact of interference on a radio link is similar to that of noise. Especially, the basic principles discussed above apply also to a scenario where interference is the main radio-link impairment:

- The maximum data rate that can be achieved in a given bandwidth is limited by the available signal-power-to-interference-power ratio.
- Providing data rates larger than the available bandwidth (high-bandwidth utilization) is costly in the sense that it requires an un-proportionally high signal-to-interference ratio.

Also similar to a scenario where noise is the dominating radio-link impairment, reducing the cell size as well as the use of multi-antenna techniques are key means to increase the achievable data rates also in an interference-limited scenario:

- Reducing the cell size will obviously reduce the number of users, and thus also the overall traffic, per cell. This will reduce the relative interference level and thus allow for higher data rates.
- Similar to the increase in signal-to-noise ratio, proper combining of the signals received at multiple antennas will also increase the signal-to-interference ratio after the antenna combining.
- The use of beam-forming by means of multiple transmit antennas will focus the transmit power in the direction of the target receiver, leading to reduced interference to other radio links and thus improving the overall signal-to-interference ratio in the system.

One important difference between interference and noise is that interference, in contrast to noise, typically has a certain structure which makes it, at least to some extent, predictable and thus possible to further suppress or even remove completely. As an example, a dominant interfering signal may arrive from a certain direction in which case the corresponding interference can be further suppressed, or even completely removed, by means of *spatial processing* using multiple antennas at the receiver. Also any differences in the spectrum properties between the target signal and an interfering signal can be used to suppress the interferer and thus reduce the overall interference level.

4.3.2 Higher data rates within a limited bandwidth: Higher-order modulation

As discussed in the previous section, providing data rates larger than the available bandwidth is fundamentally in-efficient in the sense that it requires un-proportionally high signal-to-noise and signal-to-interference ratios at the receiver. Still, bandwidth is often a scarce and expensive resource and, at least in some mobile-communication scenarios, high signal-to-noise and signal-to-interference ratios can be made available, e.g. in small-cell environments with a low traffic load or for mobile terminals close to the cell site. Future mobile-communication systems, including the evolution of 3G mobile communication, should be designed to be able to take advantage of such scenarios, that is should be able for offer very high data rates within a limited bandwidth when the radio conditions so allow.

A straightforward means to provide higher data rates within a given transmission bandwidth is the use of *higher-order modulation*, implying that the modulation alphabet is extended to include additional signaling alternatives and thus allowing for more bits of information to be communicated per modulation symbol.

In case of QPSK modulation, i.e. the modulation scheme used for the downlink in the first releases of the 3G mobile-communication standards (WCDMA and cdma2000), the modulation alphabet consists of four different signaling alternatives. These four signaling alternatives can be illustrated as four different points in a two-dimensional plane (see Figure 4.3-2a). With four different signaling alternatives, QPSK allows for up to 2 bits of information to be communicated during each modulation-symbol interval. By extending to 16QAM modulation (Figure 4.3-2b), 16 different signaling alternatives are available. The use of 16QAM thus allows for up to 4 bits of information to be communicated per symbol interval. Further extension to 64QAM (Figure 4.3-2c), with 64 different signaling alternatives, allows for up to 6 bits of information to be communicated per symbol interval. At the same time, the bandwidth of the transmitted signal is, at least in principle, independent of the size of the modulation alphabet and mainly depends on the modulation rate, i.e. the number of modulation symbols per second. The maximum bandwidth utilization, expressed in bits/s/Hz, of 16QAM and 64QAM are thus, at least in principle, two and three times that of QPSK, respectively.

It should be pointed out that there are many other possible modulation schemes, in addition to those illustrated in Figure 4.3-2. One example is 8PSK consisting of eight signaling alternatives and thus providing up to 3 bits of information per modulation symbol.

The use of higher-order modulation provides the possibility for higher bandwidth utilization, that is the possibility to provide higher data rates within a given bandwidth. However, the higher bandwidth utilization comes at the cost of reduced robustness to noise and interference. Alternatively expressed, higher-order modulation schemes, such as 16QAM or 64QAM, require a higher E_b/N_0 at the receiver for a given bit-error probability, compared to QPSK. This is in line with the discussion in the previous section where it was concluded that high bandwidth utilization, i.e. a high information rate within a limited bandwidth, in general requires a higher receiver E_b/N_0.

4.3.2.1 Higher-order modulation in combination with channel coding

Higher-order modulation schemes such as 16QAM and 64QAM require, in themselves, a higher receiver E_b/N_0 for a given error rate, compared to QPSK. However, in combination with channel coding the use of higher-order modulation will sometimes be more efficient, that is require a lower receiver E_b/N_0 for a given error rate, compared to the use of lower-order modulation such as QPSK. This may, for example, occur when the target bandwidth utilization implies that, with lower-order modulation, no or very little channel coding can be applied. In such a case, the additional channel coding that can be applied by using a higher-order modulation scheme such as 16QAM may lead to an overall gain in power efficiency compared to the use of QPSK.

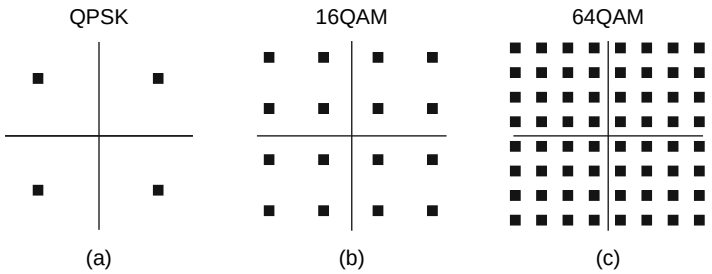


Figure 4.3-2 Signal constellations for (a) QPSK, (b) 16QAM and (c) 64QAM.

As an example, if a bandwidth utilization of close to two information bits per modulation symbol is required, QPSK modulation would allow for very limited channel coding (channel-coding rate close to one). On the other hand, the use of 16QAM modulation would allow for a channel-coding rate in the order of one half. Similarly, if a bandwidth efficiency close to 4 information bits per modulation symbol is required, the use of 64QAM may be more efficient than 16QAM modulation, taking into account the possibility for lower-rate channel coding and corresponding additional coding gain in case of 64QAM. It should be noted that this does not speak against the general discussion in Section 4.3.1 where it was concluded that transmission with high-bandwidth utilization is inherently power inefficient. The use of rate 1/2 channel coding for 16QAM obviously reduces the information data rate, and thus also the bandwidth utilization, to the same level as uncoded QPSK.

From the discussion above it can be concluded that, for a given signal-to-noise/interference ratio, a certain combination of modulation scheme and channel-coding rate is optimal in the sense that it can deliver the highest-bandwidth utilization (the highest data rate within a given bandwidth) for that signal-to-noise/interference ratio.

4.3.2.2 Variations in instantaneous transmit power

A general drawback of higher-order modulation schemes such as 16QAM and 64QAM, where information is also encoded in the instantaneous amplitude of the modulated signal, is that the modulated signal will have larger variations, and thus also larger peaks, in its instantaneous power. This can be seen from Figure 4.3-3 which illustrates the distribution of the instantaneous power, more specifically the probability that the instantaneous power is above a certain value, for QPSK, 16QAM, and 64QAM, respectively. Clearly, the probability for large peaks in the instantaneous power is higher in case of higher-order modulation.

Larger peaks in the instantaneous signal power imply that the transmitter power amplifier must be over-dimensioned to avoid the power-amplifier non-linearities that, occur at high instantaneous power levels, and cause corruption to the signal to be transmitted. As a consequence, the power-amplifier efficiency will be reduced, leading to increased power consumption. In addition, there will be a negative impact on the power-amplifier cost. Alternatively, the average transmit power must be reduced, implying a reduced range for a given data rate. High power-amplifier efficiency is especially important for the mobile terminal, i.e. in the uplink direction, due to the importance of low mobile-terminal power consumption and cost. For the base station, high power-amplifier efficiency, although far from irrelevant, is still somewhat less important. Thus, large peaks in the instantaneous signal power is less of an issue for the downlink compared to the uplink and, consequently, higher-order modulation is more suitable for the downlink compared to the uplink.

4.3.3 Wider bandwidth including multi-carrier transmission

As was shown in Section 4.3.1, transmission with a high-bandwidth utilization is fundamentally power inefficient in the sense that it will require unproportionally high signal-to-noise and signal-to-interference ratios for a given data rate. Providing very high data rates within a limited bandwidth, for example by means of higher-order modulation, is thus only possible in situations where

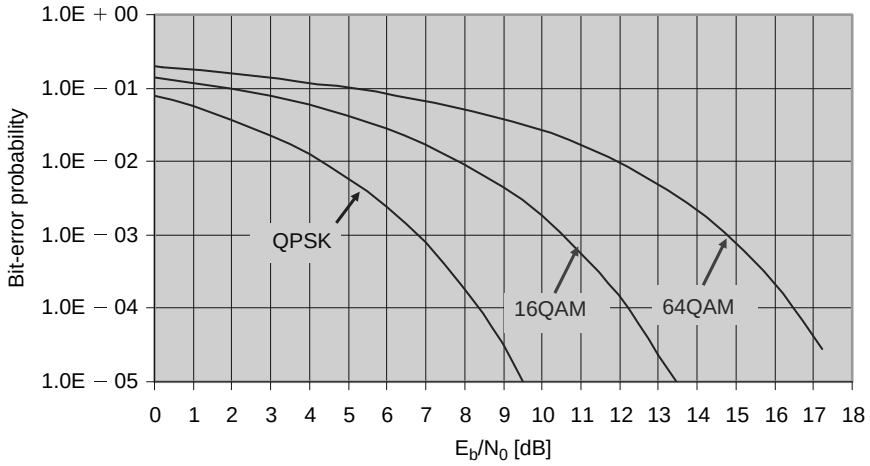


Figure 4.3-3 Distribution of instantaneous power for different modulation schemes. Same average power in all cases.

relatively high signal-to-noise and signal-to-interference ratios can be made available, for example in small-cell environments with low traffic load or for mobile terminals close to the cell site.

Instead, to provide high data rates as efficiently as possible in terms of required signal-to-noise and signal-to-interference ratios, implying as good coverage as possible for high data rates, the transmission bandwidth should be at least of the same order as the data rates to be provided.

Having in mind that the provisioning of higher data rates with good coverage is one of the main targets for the evolution of 3G mobile communication, it can thus be concluded that support for even wider transmission bandwidth is an important part of this evolution.

However, there are several critical issues related to the use of wider transmission bandwidths in a mobile-communication system:

- Spectrum is, as already mentioned, often a scarce and expensive resource, and it may be difficult to find spectrum allocations of sufficient size to allow for very wideband transmission, especially at lower-frequency bands.
- The use of wider transmission and reception bandwidths has an impact on the complexity of the radio equipment, both at the base station and at the mobile terminal. As an example, a wider transmission bandwidth has a direct impact on the transmitter and the receiver sampling rates, and thus on the complexity and power consumption of digital-to-analog and analog-to-digital converters as well as front-end digital signal processing. RF components are also, in general, more complicated to design and more expensive to produce, the wider the bandwidth they are to handle.

The two issues above are mainly outside the scope of this book. However, a more specific technical issue related to wider-band transmission is the increased corruption of the transmitted signal due to time dispersion on the radio channel. Time dispersion occurs when the transmitted signal propagates to the receiver via multiple paths with different delays (see Figure 4.3-4a). In the frequency domain, a time-dispersive channel corresponds to a non-constant channel frequency response as illustrated in Figure 4.3-4b. This radio-channel *frequency selectivity* will corrupt the frequency-domain structure of the transmitted signal and lead to higher error rates for given signal-to-noise/interference ratios. Every radio channel is subject to frequency selectivity, at least to some extent. However, the extent to which the frequency selectivity impacts the radio communication depends on the bandwidth of the transmitted signal with, in general, larger impact for wider-band transmission. The amount of radio-channel frequency selectivity also depends on the environment with typically less frequency selectivity (less time dispersion) in case of small cells and in environments with few obstructions and potential reflectors, such as rural environments.

It should be noted that Figure 4.3-4b illustrates a 'snapshot' of the channel frequency response. As a mobile terminal is moving through the environment, the detailed structure of the multi-path propagation, and thus also the detailed structure of the channel frequency response, may vary rapidly in time. The rate of the variations in the channel frequency response is related to the channel *Doppler spread*, f_D, defined as $f_D = v/c \cdot f_c$, where v is the speed of the mobile terminal, f_c is the carrier frequency (e.g. 2 GHz), and c is the speed of light.

Receiver-side *equalization* has for many years been used to counteract signal corruption due to radio-channel frequency selectivity. Equalization has been shown to provide satisfactory performance with reasonable complexity at least up to bandwidths corresponding to the WCDMA bandwidth of 5 MHz. However, if the transmission bandwidth is further increased up to, for example 20 MHz, which is the target for the 3GPP LTE, the complexity of straightforward high-performance equalization starts to become a serious issue. One option is then to apply less optimal equalization, with a corresponding negative impact on the equalizer capability to counteract the signal corruption due to radio-channel frequency selectivity and thus a corresponding negative impact on the radio-link performance.

An alternative approach is to consider specific transmission schemes and signal designs that allow for

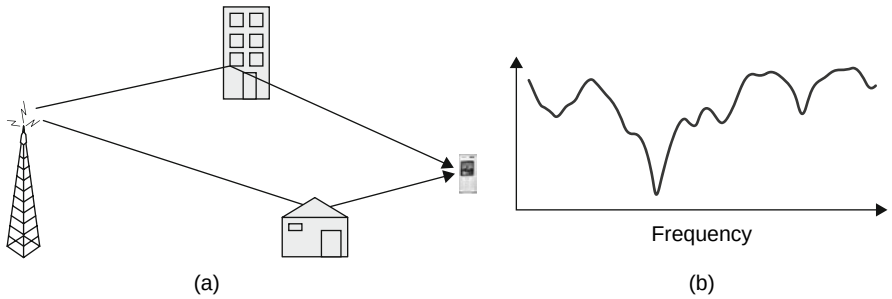


Figure 4.3-4 Multi-path propagation causing time dispersion and radio-channel frequency selectivity.

good radio-link performance in case of substantial radio-channel frequency selectivity without a prohibitively large receiver complexity. In the following, two such approaches to wider-band transmission will be discussed:

1. The use of different types of *multi-carrier transmission*, that is transmitting an overall wider-band signal as several more narrowband frequency-multiplexed signals, see below. One special case of multi-carrier transmission is *OFDM transmission* to be discussed in more detail in Chapter 4.4.
2. The use of specific *single-carrier* transmission schemes, especially designed to allow for efficient but still reasonably low-complexity equalization.

4.3.3.1 Multi-carrier transmission

One way to increase the overall transmission bandwidth, without suffering from increased signal corruption due to radio-channel frequency selectivity, is the use of so-called *multi-carrier transmission*. As illustrated in Figure 4.3-5, multi-carrier transmission implies that, instead of transmitting a single more wideband signal, multiple more narrowband signals, often referred to as *subcarriers*, are frequency multiplexed and jointly transmitted over the same radio link to the same receiver. By transmitting M signals in parallel over the same radio link, the overall data rate can be increased up to M times. At the same time, the impact in terms of signal corruption due to radio-channel frequency selectivity depends on the bandwidth of each subcarrier. Thus, the impact from a frequency-selective channel is essentially the same as for a more narrowband transmission scheme with a bandwidth that corresponds to the bandwidth of each subcarrier.

A drawback of the kind of multi-carrier evolution outlined in Figure 4.3-5, where an existing more narrowband radio-access technology is extended to a wider overall transmission bandwidth by the parallel transmission of M more narrowband carriers, is that the spectrum of each subcarrier does typically not allow for very tight subcarrier 'packing.' This is illustrated by the 'valleys' in the overall multi-carrier spectrum outlined in the lower part of Figure 4.3-5. This has a somewhat negative impact on the overall bandwidth efficiency of this kind of multi-carrier transmission.

As an example, consider a WCDMA multi-carrier evolution towards wider bandwidth. WCDMA has a modulation rate, also referred to as the *WCDMA chip rate*, of 310 $f_{cr} = 3.84$ Mchips/s. However, due to spectrum shaping, even the theoretical WCDMA spectrum, not including spectrum widening due to transmitter imperfections, has a bandwidth that significantly exceeds 3.84 MHz. More specifically, as can be seen in Figure 4.3-6, the theoretical WCDMA spectrum has a *raised-cosine* shape with roll-off $\alpha = 0.22$. As a consequence, the bandwidth outside of which the WCDMA theoretical spectrum equals zero is approximately 4.7 MHz (see right part of Figure 4.3-6).

For a straightforward multi-carrier extension of WCDMA, the subcarriers must thus be spaced

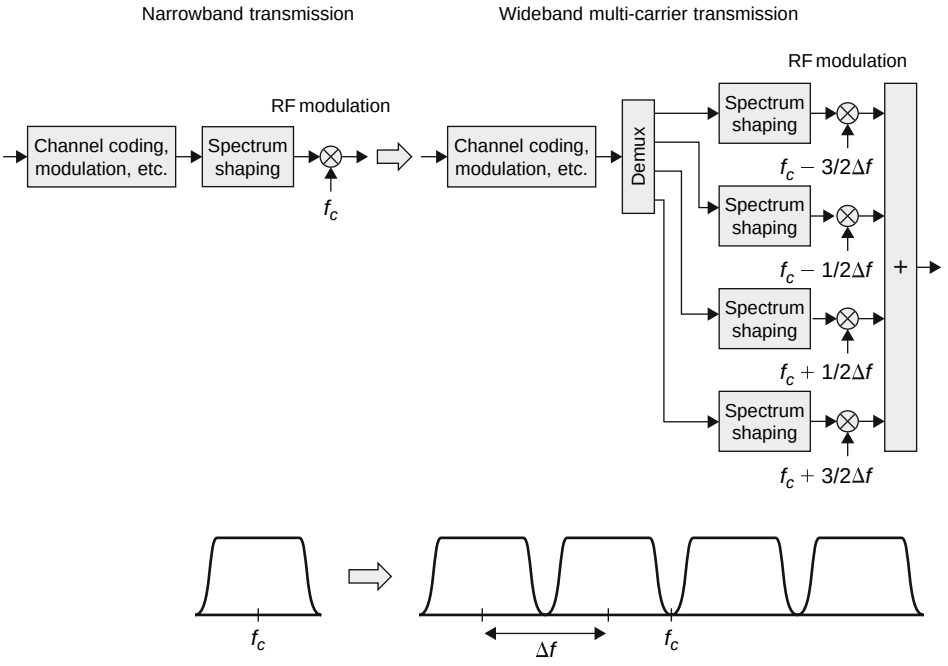
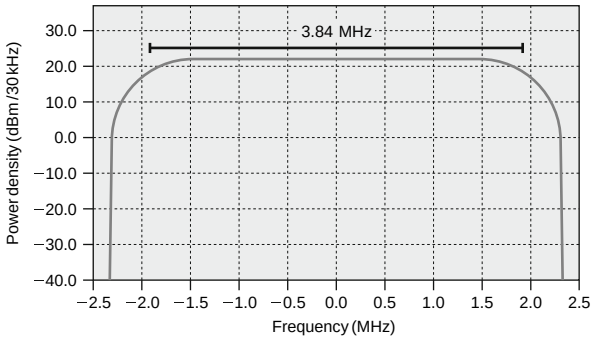


Figure 4.3-5 Extension to wider transmission bandwidth by means of multi-carrier transmission.



$$p(f) = \begin{cases} 1 & |f| < f_1 \\ \frac{1}{2}\left(1 + \cos\left(\frac{\pi(|f| - f_1)}{f_{cr}\,\alpha}\right)\right) & f_1 \le |f| < f_2 \\ 0 & |f| > f_2 \end{cases}$$

$$f_1 = f_{cr} \cdot \frac{(1-\alpha)}{2} \quad f_2 = f_{cr} \cdot \frac{(1+\alpha)}{2}$$

$$BW = 2 \cdot f_2 = f_{cr} \cdot (1-\alpha) \approx 4.7\,\text{MHz}$$

Figure 4.3-6 Theoretical WCDMA spectrum. Raised-cosine shape with roll-off $\alpha = 0.22$.

approximately 4.7 MHz from each other to completely avoid inter-subcarrier interference. It should be noted though that a smaller subcarrier spacing can be used with only limited inter-subcarrier interference.

A second drawback of multi-carrier transmission is that, similar to the use of higher-order modulation, the parallel transmission of multiple carriers will lead to larger variations in the instantaneous transmit power. Thus, similar to the use of higher-order modulation, multi-carrier transmission will have a negative impact on the transmitter power-amplifier efficiency, implying increased transmitter power consumption and increased power-amplifier cost. Alternatively, the average transmit power must be reduced, implying a reduced range for a given data rate. For this reason, similar to the use of higher-order modulation, multi-carrier transmission is more suitable for the downlink (base-station transmission), compared to the uplink (mobile-terminal transmission), due to the higher importance of high power-amplifier efficiency at the mobile terminal.

The main advantage with the kind of multi-carrier extension outlined in Figure 4.3-5 is that it provides a very smooth evolution, in terms of both radio equipment and spectrum, of an already existing radio-access technology to wider transmission bandwidth and a corresponding possibility for higher data rates, especially for the downlink. In essence this kind of multi-carrier evolution to wider bandwidth can be designed so that, for legacy terminals not capable of multi-carrier reception, each downlink 'subcarrier' will appear as an original, more narrowband carrier, while, for a multi-carrier-capable terminal, the network can make use of the full multi-carrier bandwidth to provide higher data rates.

The next chapter will discuss, in more detail, a different approach to multi-carrier transmission, based on so-called OFDM technique.

Chapter 4.4

4.4

OFDM transmission

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

In this chapter, a more detailed overview of OFDM or *Orthogonal Frequency Division Multiplexing* will be given. OFDM has been adopted as the downlink transmission scheme for the 3GPP Long-Term Evolution (LTE) and is also used for several other radio technologies, e.g. WiMAX and the DVB broadcast technologies.

4.4.1 Basic principles of OFDM

Transmission by means of OFDM can be seen as a kind of multi-carrier transmission. The basic characteristics of OFDM transmission, which distinguish it from a straightforward multi-carrier extension of a more narrowband transmission scheme as outlined in Figure 4.3-5 in the previous chapter, are:

- The use of a relatively large number of narrowband subcarriers. In contrast, a straightforward multi-carrier extension as outlined in Figure 4.3-5 would typically consist of only a few subcarriers, each with a relatively wide bandwidth. As an example, a WCDMA multi-carrier evolution to a 20 MHz overall transmission bandwidth could consist of four (sub)carriers, each with a bandwidth in the order of 5 MHz. In comparison, OFDM transmission may imply that several hundred subcarriers are transmitted over the same radio link to the same receiver.
- Simple rectangular pulse shaping as illustrated in Figure 4.4-1a. This corresponds to a sinc-square-shaped per-subcarrier spectrum, as illustrated in Figure 4.4-1b.
- Tight frequency-domain packing of the subcarriers with a subcarrier spacing $\Delta f = 1/T_u$, where T_u is the per-subcarrier modulation-symbol time (see Figure 4.4-2). The subcarrier spacing is thus equal to the per-subcarrier modulation rate $1/T_u$.

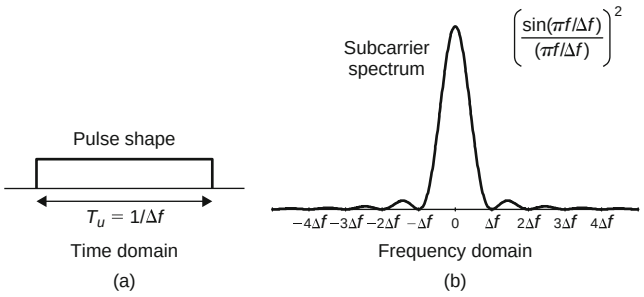


Figure 4.4-1 Per-subcarrier pulse shape and spectrum for basic OFDM transmission.

An illustrative description of a basic OFDM modulator is provided in Figure 4.4-3. It consists of a bank of N_c complex modulators, where each modulator corresponds to one OFDM subcarrier.

In complex baseband notation, a basic OFDM signal $x(t)$ during the time interval $mT_u \le t < (m+1)T_u$ can thus be expressed as

$$x(t) = \sum_{k=0}^{N_c-1} x_k(t) = \sum_{k=0}^{N_c-1} a_k^{(m)} e^{j2\pi k\Delta f t} \qquad (4.4.1)$$

where $x_k(t)$ is the kth modulated subcarrier with frequency $f_k = k \cdot \Delta f$ and $a_k^{(m)}$, in general complex, is the modulation symbol applied to the kth subcarrier during the mth OFDM symbol interval, i.e. during the time interval $mT_u \le t < (m+1)T_u$. OFDM transmission is thus block based, implying that, during each OFDM symbol interval, N_c modulation symbols are transmitted in parallel. The modulation symbols can be from any modulation alphabet, such as QPSK, 16QAM, or 64QAM.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved

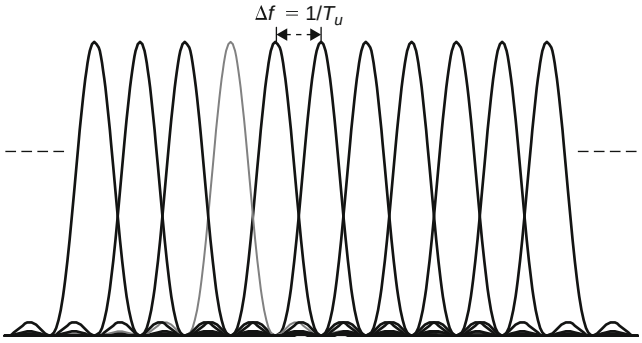


Figure 4.4-2 OFDM subcarrier spacing.

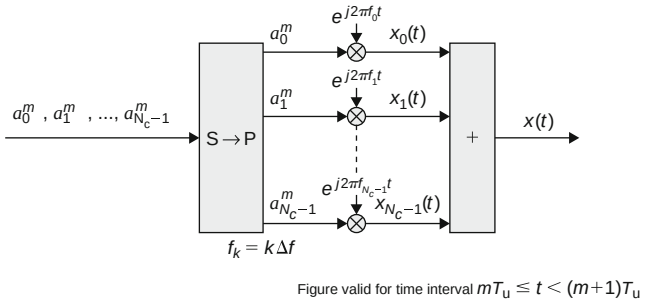


Figure 4.4-3 OFDM modulation.

The number of OFDM subcarriers can range from less than one hundred to several thousand, with the subcarrier spacing ranging from several hundred kHz down to a few kHz. What subcarrier spacing to use depends on what types of environments the system is to operate in, including such aspects as the maximum expected radio-channel frequency selectivity (maximum expected time dispersion) and the maximum expected rate of channel variations (maximum expected Doppler spread). Once the subcarrier spacing has been selected, the number of subcarriers can be decided based on the assumed overall transmission bandwidth, taking into account acceptable out-of-band emission, etc. The selection of OFDM subcarrier spacing and number of subcarriers is discussed in somewhat more detail in Section 4.4.8.

As an example, for 3GPP LTE the basic subcarrier spacing equals 15 kHz. On the other hand, the number of subcarriers depends on the transmission bandwidth, with in the order of 600 subcarriers in case of operation in a 10 MHz spectrum allocation and correspondingly fewer/more subcarriers in case of smaller/larger overall transmission bandwidths.

The term *Orthogonal Frequency Division Multiplex* is due to the fact that two modulated OFDM subcarriers $x_{k_1}(t)$ and $x_{k_2}(t)$ are mutually *orthogonal* over the time interval $mT_u \le t < (m+1)T_u$, i.e.

$$\int_{mT_u}^{(m+1)T_u} x_{k_1}(t)x_{k_2}^*(t)dt = \int_{mTu}^{(m+1)T_u} a_{k_1}a_{k_2}^* e^{j2\pi k_1 \Delta f t} e^{-j2\pi k_2 \Delta f t}\, dt = 0 \quad \text{for } k_1 \neq k_2 \tag{4.4.2}$$

Thus basic OFDM transmission can be seen as the modulation of a set of orthogonal functions $\varphi_k(t)$, where

$$\varphi_k(t) = \begin{cases} e^{j2\pi k \Delta f t} & 0 \le t < T_u \\ 0 & \text{otherwise} \end{cases} \tag{4.4.3}$$

The 'physical resource' in case of OFDM transmission is often illustrated as a time–frequency grid according to Figure 4.4-4 where each 'column' corresponds to one OFDM symbol and each 'row' corresponds to one OFDM subcarrier.

4.4.2 OFDM demodulation

Figure 4.4-5 illustrates the basic principle of OFDM demodulation consisting of a bank of correlators, one for each subcarrier. Taking into account the orthogonality between subcarriers according to (4.4.2), it is clear that, in the ideal case, two OFDM subcarriers do not cause any interference to each other after demodulation. Note that

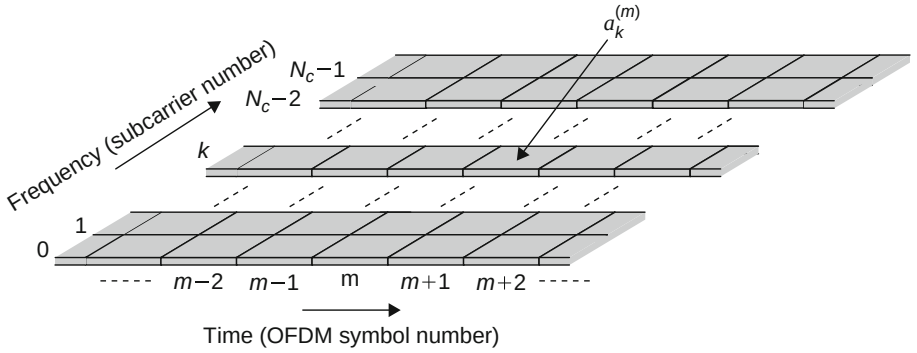


Figure 4.4-4 OFDM time–frequency grid.

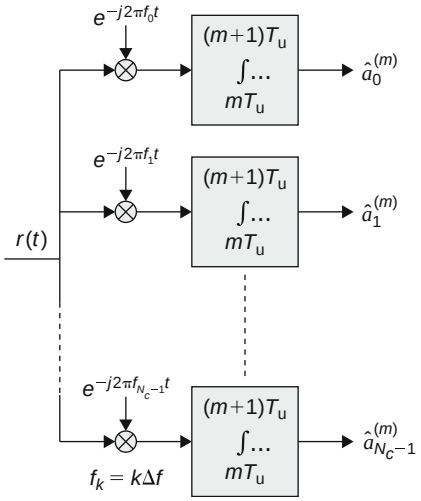


Figure 4.4-5 Basic principle of OFDM demodulation.

this is the case despite the fact that the spectrum of neighbor subcarriers clearly overlap, as can be seen from Figure 4.4-2. Thus the avoidance of interference between OFDM subcarriers is not simply due to a subcarrier spectrum separation, which is, for example, the case for the kind of straightforward multi-carrier extension outlined in Figure 4.3-5 in the previous chapter. Rather, the subcarrier orthogonality is due to the *specific* frequency-domain structure of each subcarrier in combination with the *specific* choice of a subcarrier spacing Δf equal to the per-subcarrier symbol rate $1/T_u$. However, this also implies that, in contrast to the kind of multi-carrier transmission outlined in Section 4.3.3.1 of the previous chapter, any corruption of the frequency-domain structure of the OFDM subcarriers, e.g. due to a frequency-selective radio channel, may lead to a loss of inter-subcarrier orthogonality and thus to interference between subcarriers. To handle this and to make an OFDM signal truly robust to radio-channel frequency selectivity, *cyclic-prefix insertion* is typically used, as will be further discussed in Section 4.4.4.

4.4.3 OFDM implementation using IFFT/FFT processing

Although a bank of modulators/correlators according to Figures 4.4-3 and 4.4-5 can be used to illustrate the basic principles of OFDM modulation and demodulation, respectively, these are not the most appropriate modulator/demodulator structures for actual implementation. Actually, due to its specific structure and the selection of a subcarrier spacing Δf equal to the per-subcarrier symbol rate $1/T_u$, OFDM allows for low-complexity implementation by means of computationally efficient *Fast Fourier Transform* (FFT) processing.

To confirm this, consider a time-discrete (sampled) OFDM signal where it is assumed that the sampling rate f_s is a multiple of the subcarrier spacing Δf, i.e. $f_s = 1/T_s = N \cdot \Delta f$. The parameter N should be chosen so that the sampling theorem is sufficiently fulfilled.[1] As $N_c \cdot \Delta f$ can be seen as the nominal bandwidth of the OFDM signal, this implies that N should exceed N_c with a sufficient margin.

With these assumptions, the time-discrete OFDM signal can be expressed as[2]

$$x_n = x(nT_s) = \sum_{k=0}^{N_c-1} a_k e^{j2\pi k\Delta f n T_s} = \sum_{k=0}^{N_c-1} a_k e^{j2\pi kn/N} = \sum_{k=0}^{N-1} a'_k e^{j2\pi kn/N} \quad (4.4.4)$$

where

$$a'_k = \begin{cases} a_k & 0 \le k < N_c \\ 0 & N_c \le k < N \end{cases} \quad (4.4.5)$$

Thus, the sequence x_n, i.e. the sampled OFDM signal, is the size-N Inverse Discrete Fourier Transform (IDFT) of the block of modulation symbols $a_0, a_1, \ldots, a_{N_c-1}$ extended with zeros to length N. OFDM modulation can thus be implemented by means of IDFT processing followed by digital-to-analog conversion, as illustrated in Figure 4.4-6. Especially, by selecting the IDFT size N equal to 2^m for some integer m, the OFDM modulation can be implemented by means of implementation-efficient radix-2 Inverse Fast Fourier Transform (IFFT) processing. It should be noted that the ratio N/N_c, which could be seen as the over-sampling of the time-discrete OFDM signal, could very well be, and typically is, a non-integer number. As an example and as already mentioned, for 3GPP LTE the number of subcarriers N_c is approximately 600 in case of a 10 MHz spectrum allocation. The IFFT size can then, for example, be selected as $N = 1024$. This corresponds to a sampling rate $f_s = N \cdot \Delta f = 15.36$ MHz, where $\Delta f = 15$ kHz is the LTE subcarrier spacing.

It is important to understand that IDFT/IFFT-based implementation of an OFDM modulator, and even more so the exact IDFT/IFFT size, are just transmitter-implementation choices and not something that would

[1] An OFDM signal defined according to (4.4.1) in theory has an infinite bandwidth and thus the sampling theorem can never be fulfilled completely.
[2] From now on the index m on the modulation symbols, indicating the OFDM-symbol number, will be ignored unless especially needed.

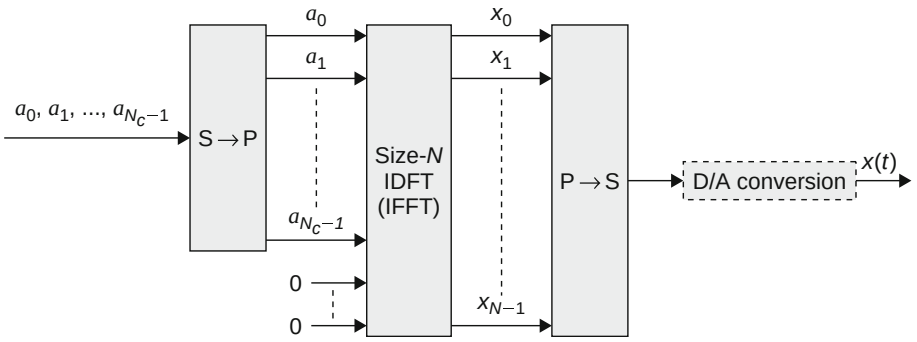


Figure 4.4-6 OFDM modulation by means of IFFT processing.

be mandated by any radio-access specification. As an example, nothing forbids the implementation of an OFDM modulator as a set of parallel modulators as illustrated in Figure 4.4-3. Also nothing prevents the use of a larger IFFT size, e.g. a size-2048 IFFT size, even in case of a smaller number of OFDM subcarriers.

Similar to OFDM modulation, efficient FFT processing can be used for OFDM demodulation, replacing the bank of N_c parallel demodulators of Figure 4.4-5 with sampling with some sampling rate $f_s = 1/T_s$, followed by a size-N DFT/FFT, as illustrated in Figure 4.4-7.

4.4.4 Cyclic-prefix insertion

As described in Section 4.4.2, an uncorrupted OFDM signal can be demodulated without any interference between subcarriers. One way to understand this *subcarrier orthogonality* is to recognize that a modulated subcarrier $x_k(t)$ in (4.4.1) consists of an integer number of periods of complex exponentials during the demodulator integration interval $T_u = 1/\Delta f$.

However, in case of a time-dispersive channel the orthogonality between the subcarriers will, at least partly, be lost. The reason for this loss of subcarrier orthogonality in case of a time-dispersive channel is that, in this case, the demodulator correlation interval for one path will overlap with the symbol boundary of a different path, as illustrated in Figure 4.4-8. Thus, the integration interval will not necessarily correspond to an integer number of periods of complex exponentials of that path as the modulation symbols a_k may differ between consecutive symbol intervals. As a consequence, in case of a time-dispersive channel there will not only be inter-symbol interference within a subcarrier but also interference between subcarriers.

Another way to explain the interference between subcarriers in case of a time-dispersive channel is to have in mind that time dispersion on the radio channel is equivalent to a frequency-selective channel frequency response. As clarified in Section 4.4.2, orthogonality between OFDM subcarriers is not simply due to frequency-domain separation but due to the *specific* frequency-domain structure of each subcarrier. Even if the frequency-domain channel is constant over a bandwidth corresponding to the main lobe of an OFDM subcarrier and only the subcarrier side lobes are corrupted due to the radio-channel frequency selectivity, the orthogonality between subcarriers will be lost with inter-subcarrier interference as a consequence. Due to the relatively large side lobes of each OFDM subcarrier, already a relatively limited amount of time dispersion or, equivalently, a relatively modest radio-channel frequency selectivity may cause non-negligible interference between subcarriers.

To deal with this problem and to make an OFDM signal truly insensitive to time dispersion on the radio channel, so-called *cyclic-prefix insertion* is typically used in case of OFDM transmission. As illustrated in Figure 4.4-9, cyclic-prefix insertion implies that the last part of the OFDM symbol is copied and inserted at the beginning of the OFDM symbol. Cyclic-prefix insertion thus increases the length of the OFDM symbol from T_u to $T_u + T_{CP}$ where T_{CP} is the length of the cyclic prefix, with a corresponding reduction in the OFDM symbol rate as a consequence. As illustrated in the lower part of Figure 4.4-9, if the correlation at the receiver side is still only carried out over a time interval $T_u = 1/\Delta f$, subcarrier orthogonality will then be preserved in case of

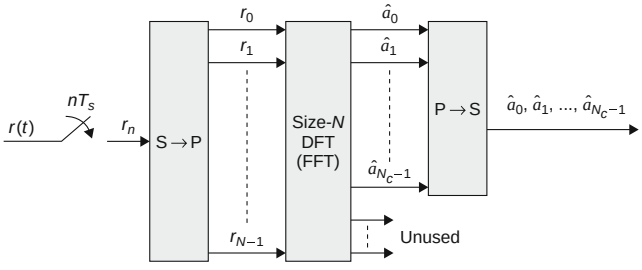


Figure 4.4-7 OFDM demodulation by means of FFT processing.

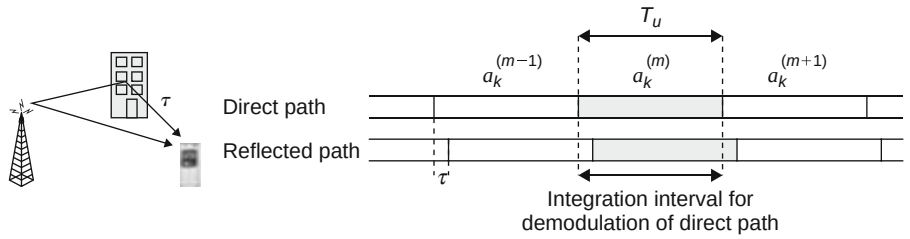


Figure 4.4-8 Time dispersion and corresponding received-signal timing.

a time-dispersive channel, as long as the span of the time dispersion is shorter than the cyclic-prefix length.

In practice, cyclic-prefix insertion is carried out on the time-discrete output of the transmitter IFFT. Cyclic-prefix insertion then implies that the last N_{CP} samples of the IFFT output block of length N is copied and inserted at the beginning of the block, increasing the block length from N to $N + N_{CP}$. At the receiver side, the corresponding samples are discarded before OFDM demodulation by means of, for example, DFT/FFT processing.

Cyclic-prefix insertion is beneficial in the sense that it makes an OFDM signal insensitive to time dispersion as long as the span of the time dispersion does not exceed the length of the cyclic prefix. The drawback of cyclic-prefix insertion is that only a fraction $T_u/(T_u + T_{CP})$ of the received signal power is actually utilized by the OFDM demodulator, implying a corresponding power loss in the demodulation. In addition to this power loss, cyclic-prefix insertion also implies a corresponding loss in terms of bandwidth as the OFDM symbol rate is reduced without a corresponding reduction in the overall signal bandwidth.

One way to reduce the relative overhead due to cyclic-prefix insertion is to reduce the subcarrier spacing Δf, with a corresponding increase in the symbol time T_u as a consequence. However, this will increase the sensitivity of the OFDM transmission to fast channel variations, that is high Doppler spread, as well as different types of frequency errors, see further Section 4.4.8.

It is also important to understand that the cyclic prefix does not necessarily have to cover the entire length of the channel time dispersion. In general, there is a trade-off between the power loss due to the cyclic prefix and the signal corruption (inter-symbol and inter-subcarrier interference) due to residual time dispersion not covered by the cyclic prefix and, at a certain point, further reduction of the signal corruption due to further increase of the cyclic-prefix length will not justify the corresponding additional power loss. This also means that, although the amount of time dispersion typically increases with the cell size, beyond a certain cell size there is often no reason to increase the cyclic prefix further as the corresponding power loss due to a further increase of the cyclic prefix would have a larger negative impact, compared to the signal corruption due to the residual time dispersion not covered by the cyclic prefix.

4.4.5 Frequency-domain model of OFDM transmission

Assuming a sufficiently large cyclic prefix, the linear convolution of a time-dispersive radio channel will

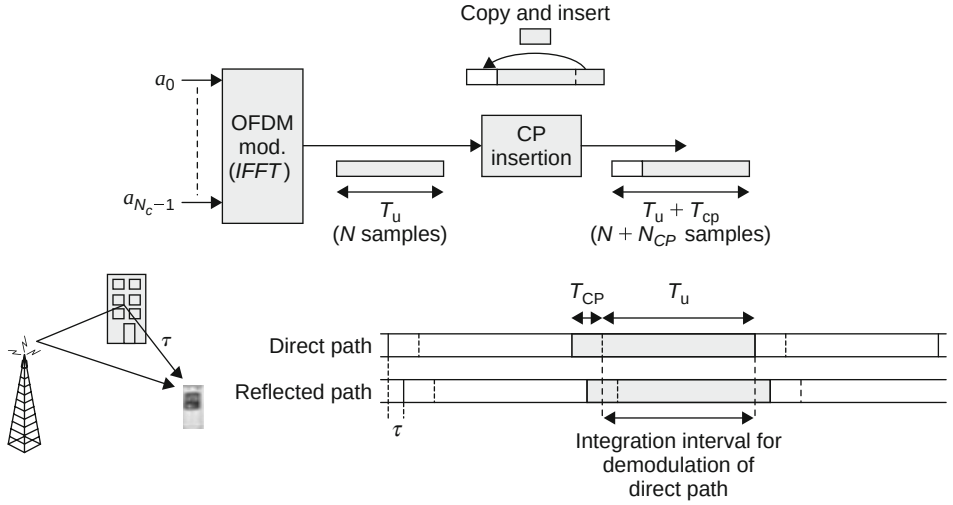


Figure 4.4-9 Cyclic-prefix insertion.

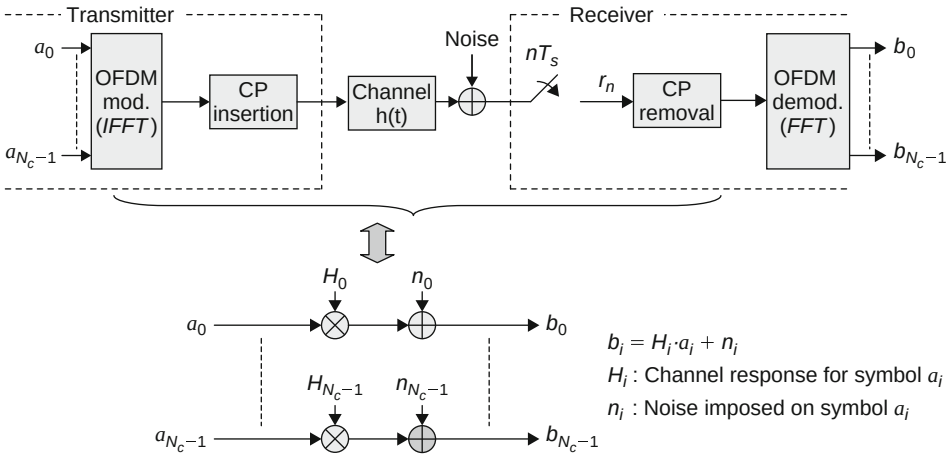


Figure 4.4-10 Frequency-domain model of OFDM transmission/reception.

appear as a circular convolution during the demodulator integration interval T_u. The combination of OFDM modulation (IFFT processing), a time-dispersive radio channel, and OFDM demodulation (*FFT* processing) can then be seen as *a frequency-domain* channel as illustrated in Figure 4.4-10, where the frequency-domain channel taps $H_0,\ldots, H_{N_{c-1}}$ can be directly derived from the channel impulse response.

The demodulator output b_k in Figure 4.4-10 is the transmitted modulation symbol a_k scaled and phase rotated by the complex frequency-domain channel tap H_k and impaired by noise n_k. To properly recover the transmitted symbol for further processing, for example data demodulation and channel decoding, the receiver should multiply b_k with the complex conjugate of H_k, as illustrated in Figure 4.4-11, This is often expressed as a *one-tap equalizer* being applied to each received subcarrier.

4.4.6 Channel estimation and reference symbols

As described above, to demodulate the transmitted modulation symbol a_k and allow for proper decoding of the transmitted information at the receiver side, scaling with the complex conjugate of the frequency-domain channel tap H_k should be applied after OFDM demodulation (*FFT* processing) (see Figure 4.4-11). To be able to do this, the receiver obviously needs an estimate of the frequency-domain channel taps $H_0,\ldots, H_{N_{c-1}}$.

The frequency-domain channel taps can be estimated indirectly by first estimating the channel impulse response and, from that, calculate an estimate of H_k. However, a more straightforward approach is to estimate the frequency-domain channel taps directly. This can be done by inserting known *reference symbols*, sometimes also referred to as *pilot symbols*, at regular intervals within the OFDM time-frequency grid, as illustrated in Figure 4.4-12. Using knowledge about the reference symbols, the receiver can estimate the frequency-domain channel around the location of the reference symbol. The reference symbols should have a sufficiently high density in both the time and the frequency domain to be able to provide estimates for the entire time/frequency grid in case of radio channels subject to high frequency and/or time selectivity.

Different algorithms can be used for the channel estimation, ranging from simple averaging in combination

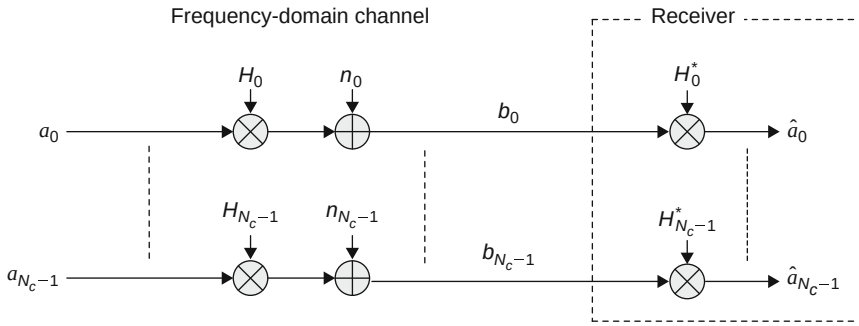


Figure 4.4-11 Frequency-domain model of OFDM transmission/reception with 'one-tap equalization' at the receiver.

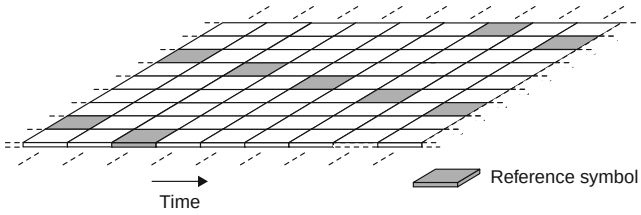


Figure 4.4-12 Time-frequency grid with known reference symbols.

with linear interpolation to Minimum Mean-Square Error (MMSE) estimation relying on more detailed knowledge of the channel time/frequency-domain characteristics.

4.4.7 Frequency diversity with OFDM: Importance of channel coding

As discussed in Section 4.3.3 in the previous chapter, a radio channel is always subject to some degree of frequency selectivity, implying that the channel quality will vary in the frequency domain. In case of a single wideband carrier, such as a WCDMA carrier, each modulation symbol is transmitted over the entire signal bandwidth. Thus, in case of the transmission of a single wideband carrier over a highly frequency-selective channel (see Figure 4.4-13a), each modulation symbol will be transmitted both over frequency bands with relatively good quality (relatively high signal strength) and frequency bands with low quality (low signal strength). Such transmission of information over multiple frequency bands with different instantaneous channel quality is also referred to *as frequency diversity*.

On the other hand, in case of OFDM transmission each modulation symbol is mainly confined to a relatively narrow bandwidth. Thus, in case of OFDM transmission over a frequency-selective channel, certain modulation symbols may be fully confined to a frequency band with very low instantaneous signal strength as illustrated in Figure 4.4-13b. Thus, the individual modulation symbols will typically not experience any substantial frequency diversity even if the channel is highly frequency selective over the overall OFDM transmission bandwidth. As a consequence, the basic error-rate performance of OFDM transmission over a frequency-selective channel is relatively poor and especially much worse than the basic error rate in case of a single wideband carrier.

However, in practice channel coding is used in most cases of digital communication and especially in case of mobile communication. Channel coding implies that each bit of information to be transmitted is spread over several, often very many, code bits. If these coded bits are then, via modulation symbols, mapped to a set of OFDM subcarriers that are well distributed over the overall transmission bandwidth of the OFDM signal, as illustrated in Figure 4.4-14, each information bit will experience frequency diversity in case of transmission over a radio channel that is frequency selective over the transmission bandwidth, despite the fact that the subcarriers, and thus also the code bits, will not experience any frequency diversity. Distributing the code bits in the frequency domain, as illustrated in Figure 4.4-14, is sometimes referred to as *frequency interleaving*. This is similar to the use of time-domain interleaving to benefit from channel coding in case of fading that varies in time.

Thus, in contrast to the transmission of a single wideband carrier, channel coding (combined with frequency interleaving) is an essential component in order for OFDM transmission to be able to benefit from frequency diversity on a frequency-selective channel. As channel coding is typically used in most cases of mobile communication this is not a very serious drawback, especially taking into account that a significant part of the available frequency diversity can be captured already with a relatively high code rate.

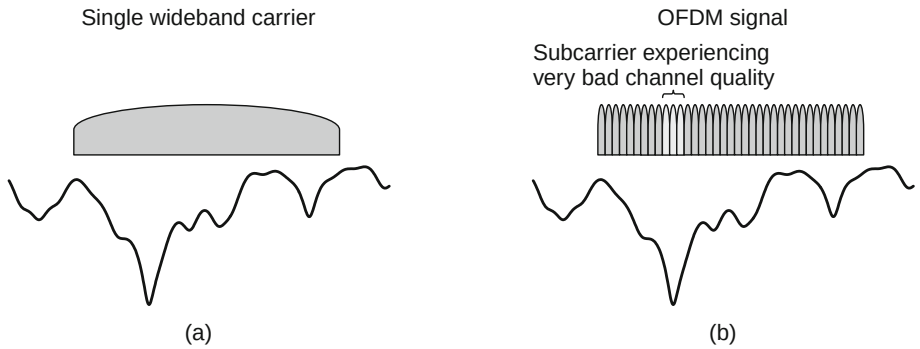


Figure 4.4-13 Transmission of single wideband carrier and OFDM transmission over a frequency-selective channel.

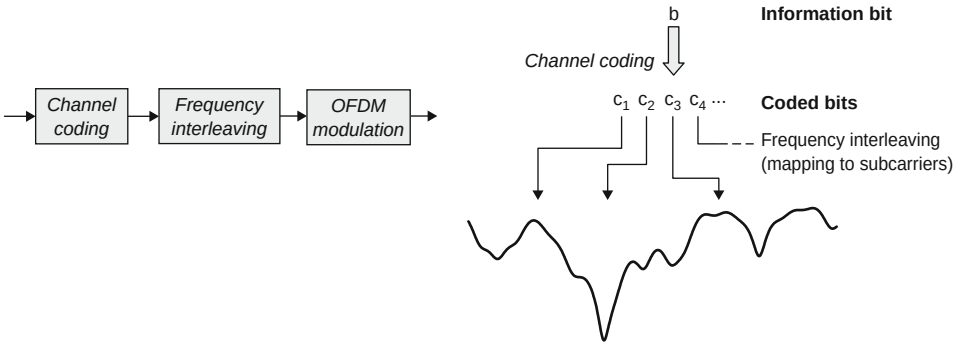


Figure 4.4-14 Channel coding in combination with frequency-domain interleaving to provide frequency diversity in case of OFDM transmission.

4.4.8 Selection of basic OFDM parameters

If OFDM is to be used as the transmission scheme in a mobile-communication system, the following basic OFDM parameters need to be decided on:

- The subcarrier spacing Δf.
- The number of subcarriers N_c, which, together with the subcarrier spacing, determines the overall transmission bandwidth of the OFDM signal.
- The cyclic-prefix length T_{CP}. Together with the subcarrier spacing $\Delta f = 1/T_u$, the cyclic-prefix length determines the overall OFDM symbol time $T = T_{CP} + T_u$ or, equivalently, the OFDM symbol rate.

4.4.8.1 OFDM subcarrier spacing

There are two factors that constrain the selection of the OFDM subcarrier spacing:

- The OFDM subcarrier spacing should be as small as possible (T_u as large as possible) to minimize the relative cyclic-prefix overhead $T_{CP}/(T_u + T_{CP})$, see further Section 4.4.8.3.
- A too small subcarrier spacing increases the sensitivity of the OFDM transmission to Doppler spread and different kinds of frequency inaccuracies.

A requirement for the OFDM subcarrier orthogonality (4.4.2) to hold at the receiver side, i.e. after the transmitted signal has propagated over the radio channel, is that the instantaneous channel does not vary noticeably during the demodulator correlation interval T_u (see Figure 4.4-5). In case of such channel variations, e.g. due very high Doppler spread, the orthogonality between subcarriers will be lost with inter-subcarrier interference as a consequence. Figure 4.4-15 illustrates the subcarrier signal-to-interference ratio due to inter-subcarrier interference between two neighbor subcarriers, as a function of the normalized Doppler spread. When considering Figure 4.4-15, it should be had in mind that a subcarrier will be subject to interference from multiple subcarriers on both sides,[3] that is the overall inter-subcarrier interference from all subcarriers will be higher than what is illustrated in Figure 4.4-15.

In practice, the amount of inter-subcarrier interference that can be accepted very much depends on the service to be provided and to what extent the received signal is corrupted due to noise and other impairments. As an example, on the cell border of large cells the signal-to-noise/interference ratio will be relatively low, with relatively low achievable data rates as a consequence. A small amount of additional inter-subcarrier interference, for example, due to Doppler spread, may then be more or less negligible. At the same time, in high signal-to-noise/interference scenarios, for example, in small cells with low traffic or close to the base station, where high data rates are to be provided, the same amount of inter-subcarrier interference may have a much more negative impact.

It should also be noted that, in addition to Doppler spread, inter-subcarrier interference will also be due to different transmitter and receiver inaccuracies, such as frequency errors and phase noise.

4.4.8.2 Number of subcarriers

Once the subcarrier spacing has been selected based on environment, expected Doppler spread and time dispersion, etc., the number of subcarriers can be determined based on the amount of spectrum available and the acceptable out-of-band emissions.

[3] Except for the subcarrier at the edge of the spectrum.

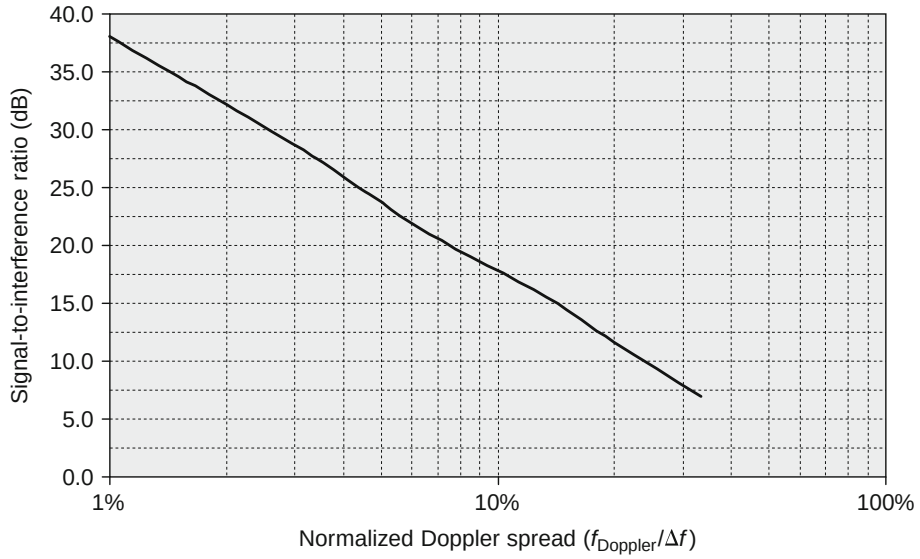


Figure 4.4-15 Subcarrier interference as a function of the normalized Doppler spread $f_{Doppler}/\Delta f$.

The basic bandwidth of an OFDM signal equals $N_c \cdot \Delta f$, i.e. the number of subcarriers multiplied by the subcarrier spacing. However, as can be seen in Figure 4.4-16, the spectrum of a basic OFDM signal falls off very slowly outside the basic OFDM bandwidth and especially much slower than for a WCDMA signal. The reason for the large *out-of-band emission* of a basic OFDM signal is the use of rectangular pulse shaping (Figure 4.4-1), leading to per-subcarrier side lobes that fall off relatively slowly. However, in practice, straightforward filtering or *time-domain windowing* will be used to suppress a main part of the OFDM out-of-band emissions. Thus, in practice, typically in the order of 10% guard-band is needed for an OFDM signal implying, as an example, that in a spectrum allocation of 5 MHz, the basic OFDM bandwidth $N_c \cdot \Delta f$ could be in the order of 4.5 MHz. Assuming, for example, a subcarrier spacing of 15 kHz as selected for LTE, this corresponds to approximately 300 subcarriers in 5 MHz.

4.4.8.3 Cyclic-prefix length

In principle, the cyclic-prefix length T_{CP} should cover the maximum length of the time-dispersion expected to be experienced. However, as already discussed, increasing the length of the cyclic prefix, without a corresponding reduction in the subcarrier spacing Δf, implies an additional overhead in terms of power as well as bandwidth. Especially the power loss implies that, as the cell size grows and the system performance becomes more power limited, there is a trade-off between the

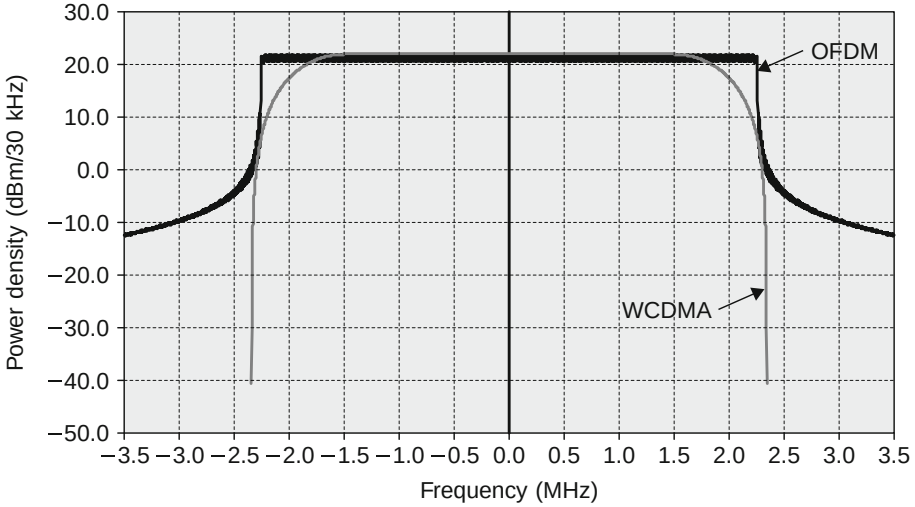


Figure 4.4-16 Spectrum of a basic 5 MHz OFDM signal compared with WCDMA spectrum.

loss in power due to the cyclic prefix and the signal corruption due to time dispersion not covered by the cyclic prefix. As already mentioned, this implies that, although the amount of time dispersion typically increases with the cell size, beyond a certain cell size there is often no reason to increase the cyclic prefix further as the corresponding power loss would have a larger negative impact, compared to the signal corruption due to the residual time dispersion not covered by the cyclic prefix.

One situation where a longer cyclic prefix may be needed is in the case of multi-cell transmission using SFN (Single-Frequency Network), as further discussed in Section 4.4.11.

Thus, to be able to optimize performance to different environments, some OFDM-based systems support multiple cyclic-prefix lengths. The different cyclic-prefix lengths can then be used in different transmission scenarios:

- Shorter cyclic prefix in small-cell environments to minimize the cyclic-prefix overhead.
- Longer cyclic prefix in environments with extreme time dispersion and especially in case of SFN operation.

4.4.9 Variations in instantaneous transmission power

According to Section 4.3.3.1, one of the drawbacks of multi-carrier transmission are the corresponding large variations in the instantaneous transmit power, implying a reduced power-amplifier efficiency and higher mobile-terminal power consumption, alternatively; the power-amplifier output power has to be reduced with a reduced range as a consequence. Being a kind of multi-carrier transmission scheme, OFDM is subject to the same drawback.

However, a large number of different methods have been proposed to reduce the large power peaks of an OFDM signal:

- In case of *tone reservation*, a subset of the OFDM subcarriers are not used for data transmission. Instead, these subcarriers are modulated in such a way that the largest peaks of the overall OFDM signal are suppressed, allowing for a reduced power-amplifier back-off. One drawback of tone reservation is the bandwidth loss due to the fact that a number of subcarriers are not available for actual data transmission. The calculation of what modulation to apply to the reserved tones can also be of relatively high complexity.
- In case of *pre-filtering* or *pre-coding*, linear processing is applied to the sequence of modulation symbols before OFDM modulation. DFT-spread-OFDM can be seen as one kind of pre-filtering.
- In case of *selective scrambling*, the coded-bit sequence to be transmitted is scrambled with a number of different scrambling codes. Each scrambled sequence is then OFDM modulated and the signal with the lowest peak power is selected for transmission. After OFDM demodulation at the receiver side, descrambling and subsequent decoding is carried out for all the possible scrambling sequences. Only the decoding carried out for the scrambling code actually used for the transmission will provide a correct decoding result. A drawback of selective scrambling is an increased receiver complexity as multiple decodings need to be carried out in parallel.

4.4.10 OFDM as a user-multiplexing and multiple-access scheme

The discussion has, until now, implicitly assumed that all OFDM subcarriers are transmitted from the same transmitter to a certain receiver, i.e.:

- downlink transmission of all subcarriers to a *single* mobile terminal;
- uplink transmission of all subcarriers from a *single* mobile terminal.

However, OFDM can also be used as a *user-multiplexing* or *multiple-access scheme*, allowing for simultaneous frequency-separated transmissions to/from multiple mobile terminals (see Figure 4.4-17).

In the downlink direction, OFDM as a user-multiplexing scheme implies that, in each OFDM symbol interval, different subsets of the overall set of available subcarriers are used for transmission to *different* mobile terminals (see Figure 4.4-17a).

Similarly, in the uplink direction, OFDM as a *user-multiplexing* or *multiple-access scheme* implies that, in each OFDM symbol interval, different subsets of the overall set of subcarriers are used for data transmission from *different* mobile terminals (see Figure 4.4-17b). In this case, the term Orthogonal Frequency Division Multiple Access or OFDMA is also often used.[4]

[4] The term OFDMA is sometimes also used to denote the use of OFDM to multiplex multiple users in the downlink as illustrated in Figure 4.4-17a.

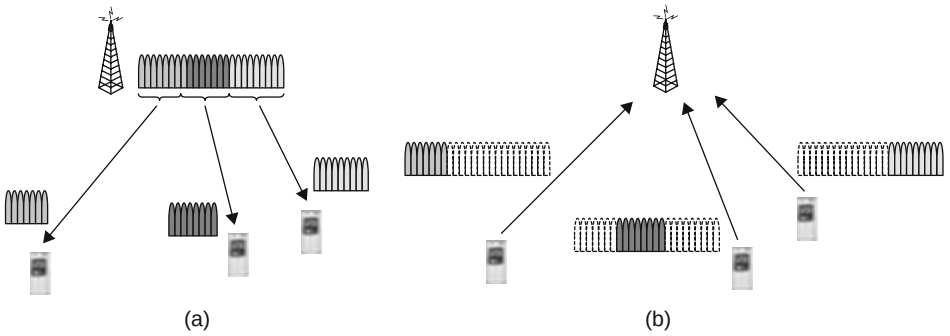


Figure 4.4-17 OFDM as a user-multiplexing/multiple-access scheme: (a) downlink and (b) uplink.

Figure 4.4-17 assumes that *consecutive* subcarriers are used for transmission to/from the same mobile terminal. However, distributing the subcarriers to/from a mobile terminal in the frequency domain is also possible as illustrated in Figure 4.4-18. The benefit of such *distributed* user multiplexing or *distributed* multiple access is a possibility for additional frequency diversity as each transmission is spread over a wider bandwidth.

In the case when OFDMA is used as an uplink multiple-access scheme, i.e. in case of frequency multiplexing of OFDM signals from multiple mobile terminals, it is critical that the transmissions from the different mobile terminals arrive approximately time aligned at the base station. More specifically, the transmissions from the different mobile terminals should arrive at the base station with a timing misalignment less than the length of the cyclic prefix to preserve orthogonality between subcarriers received from different mobile terminals and thus avoid inter-user interference.

Due to the differences in distance to the base station for different mobile terminals and the corresponding differences in the propagation time (which may far exceed the length of the cyclic prefix), it is necessary to control the uplink transmission timing of each mobile terminal (see Figure 4.4-19). Such *transmit-timing control* should adjust the transmit timing of each mobile terminal to ensure that uplink transmissions arrive approximately time aligned at the base station. Since the propagation time changes as the mobile terminal is moving within the cell, the transmit-timing control should be an active process, continuously adjusting the exact transmit timing of each mobile terminal.

Furthermore, even in case of perfect transmit-timing control, there will always be some interference between subcarriers e.g. due to frequency errors. Typically this interference is relatively low in case of reasonable frequency errors, Doppler spread, etc. (see Section 4.4.8). However, this assumes that the different subcarriers are received with at least approximately the same power. In the uplink, the propagation distance and thus the path loss of the different mobile-terminal transmissions may differ significantly. If two terminals are transmitting with the same power, the received-signal strengths may thus differ significantly, implying a potentially significant interference from the stronger signal to the weaker signal unless the subcarrier orthogonality is perfectly retained. To avoid this, at least some degree of *uplink transmit-power control* may need to be applied in case of uplink

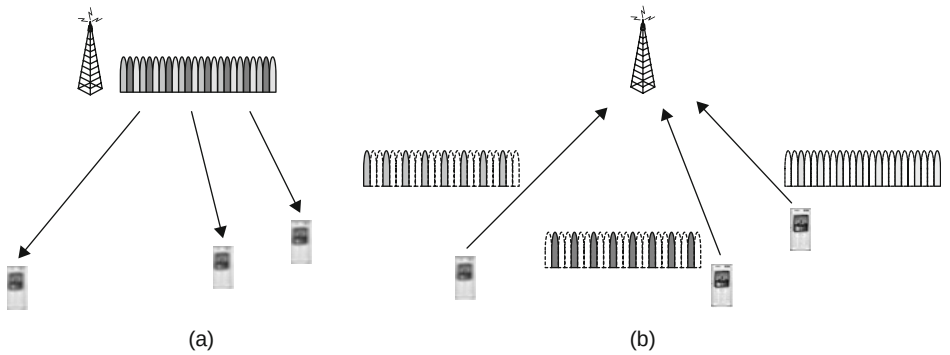


Figure 4.4-18 Distributed user multiplexing.

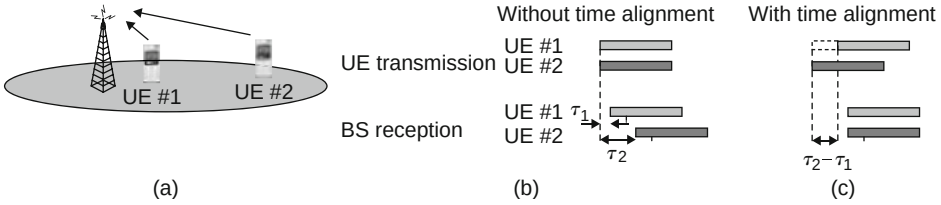


Figure 4.4-19 Uplink transmission-timing control.

OFDMA, reducing the transmit power of user terminals close to the base station and ensuring that all received signals will be of approximately the same power.

4.4.11 Multi-cell broadcast/ multicast transmission and OFDM

The provisioning of broadcast/multicast services in a mobile-communication system implies that the same information is to be *simultaneously* provided to multiple mobile terminals, often dispersed over a large area corresponding to a large number of cells as shown in Figure 4.4-20. The broadcast/multicast information may be a TV news clip, information about the local weather conditions, stock-market information, or any other kind of information that, at a given time instant, may be of interest to a large number of people.

When the same information is to be provided to multiple mobile terminals within a cell it is often beneficial to provide this information as a single 'broadcast' radio transmission covering the entire cell and simultaneously being received by all relevant mobile terminals (Figure 4.4-21a), rather than providing the information by means of individual transmissions to each mobile terminal (unicast transmission, see Figure 4.4-21b).

As a broadcast transmission according to Figure 4.4-21a has to be dimensioned to reach also the worst-case mobile terminals, including mobile terminals at the cell border, it will be relatively costly in terms of the resourses (base-station transmit power) needed to provide a certain broadcast-service data rate. Alternatively, taking into account the limited signal-to-noise ratio that can be achieved at, for example, the cell edge, the achievable broadcast data rates may be relatively limited, especially in case of large cells. One way to increase the broadcast data rates would then be to reduce the cell size, thereby increasing the cell-edge receive power. However, this will increase the number of cells to cover a certain area and is thus obviously negative from a cost-of-deployment point-of-view.

However, as discussed above, the provisioning of broadcast/multicast services in a mobile-communication network typically implies that identical information is to be provided over a large number of cells. In such a case, the resources (downlink transmit power) needed to provide a certain broadcast data rate can be considerably reduced if mobile terminals at the cell edge can utilize the received power from broadcast transmissions from multiple cells when detecting/decoding the broadcast data.

Especially large gains can be achieved if the mobile terminal can simultaneously receive and combine the

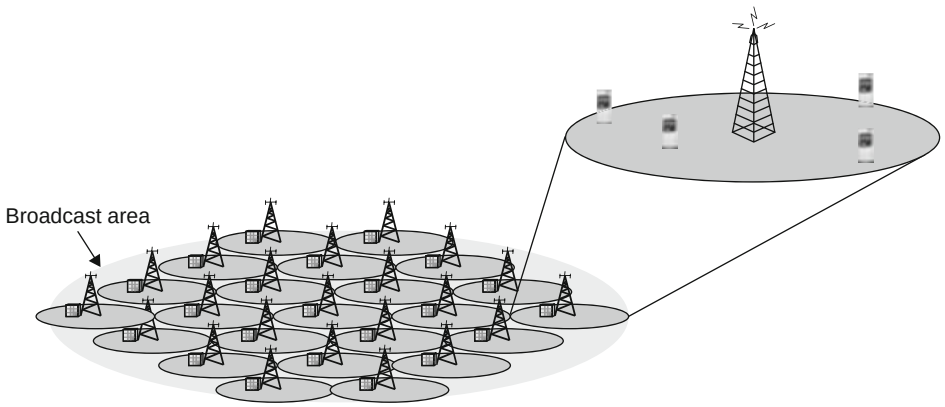


Figure 4.4-20 Broadcast scenario.

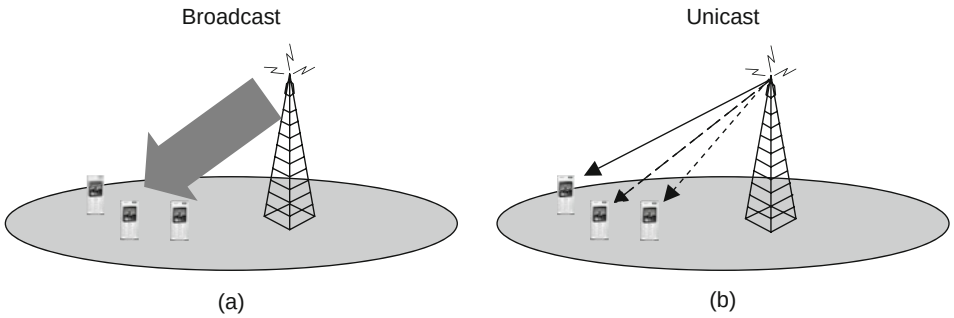


Figure 4.4-21 Broadcast vs. unicast transmission. (a) Broadcast and (b) unicast.

broadcast transmissions from multiple cells *before* decoding. Such *soft combining* of broadcast/multicast transmissions from multiple cells has already been adopted for WCDMA Multimedia Broadcast/Multicast Service (MBMS).

In case of WCDMA, each cell is transmitting on the downlink using different cell-specific scrambling codes. This is true also in case of MBMS. Thus the broadcast transmissions received from different cells are not identical but just provide the same broadcast/multicast information. As a consequence, the mobile terminal must still explicitly identify what cells to include in the soft combining. Furthermore, although the soft combining significantly increases the overall received power for mobile terminals at the cell border and thus significantly improve the overall broadcast efficiency, the broadcast transmissions from the different cells will still interfere with each other. This will limit the achievable broadcast-transmission signal-to-interference ratio and thus limit the achievable broadcast data rates.

One way to mitigate this and further improve the provisioning of broadcast/multicast services in a mobile-communication network is to ensure that the broadcast transmissions from different cells *are truly identical* and *transmitted mutually time aligned*. In this case, the transmissions received from multiple cells will, as seen from the mobile terminal, appear as a single transmission subject to severe multi-path propagation as illustrated in Figure 4.4-22. The transmission of identical time-aligned signals from multiple cells, especially in the case of provisioning of broadcast/multicast services, is sometimes referred to as Single-Frequency Network (SFN) operation.

In case of such identical time-aligned transmissions from multiple cells, the 'inter-cell interference' due to transmissions in neighboring cells will, from a terminal point of view, be replaced by signal corruption due to time dispersion. If the broadcast transmission is based on OFDM with a cyclic prefix that covers the main part of this 'time dispersion', the achievable broadcast data rates are thus only limited by noise, implying that, especially in smaller cells, very high broadcast data rates can be achieved. Furthermore, in contrast to the explicit (in the receiver) multi-cell soft combining of WCDMA MBMS, the OFDM receiver does not need to explicitly identify the cells to be soft combined. Rather, all transmissions that fall within the cyclic prefix will 'automatically' be captured by the receiver.

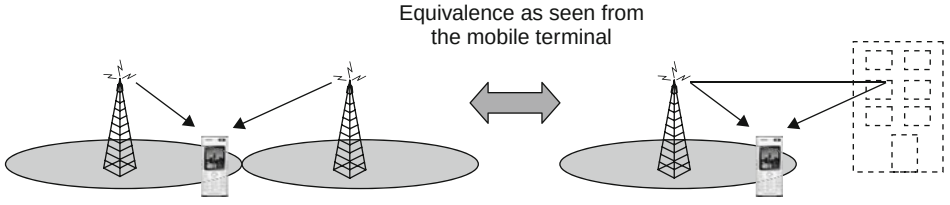


Figure 4.4-22 Equivalence between simulcast transmission and multi-path propagation.

This page intentionally left blank

Chapter 4.5

4.5

Scheduling, link adaptation and hybrid ARQ

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

One key characteristic of mobile radio communication is the typically rapid and significant variations in the instantaneous channel conditions. There are several reasons for these variations. Frequency-selective fading will result in rapid and random variations in the channel attenuation. Shadow fading and distance-dependent path loss will also affect the average received signal strength significantly. Finally, the interference at the receiver due to transmissions in other cells and by other terminals will also impact the interference level. Hence, to summarize, there will be rapid, and to some extent random, variations in the experienced quality of each radio link in a cell, variations that must be taken into account and preferably exploited.

In this chapter, some of the techniques for handling variations in the instantaneous radio-link quality will be discussed. C*hannel-dependent scheduling* in a mobile-communication system deals with the question of how to share, between different users (different mobile terminals), the radio resource(s) available in the system to achieve as efficient resource utilization as possible. Typically, this implies to minimize the amount of resources needed per user and thus allow for as many users as possible in the system, while still satisfying whatever quality-of-service requirements that may exist. Closely related to scheduling is *link adaptation*, which deals with how to set the transmission parameters of a radio link to handle variations of the radio-link quality.

Both channel-dependent scheduling and link adaptation tries to exploit the channel variations through appropriate processing *prior* to transmission of the data. However, due to the random nature of the variations in the radio-link quality, perfect adaptation to the instantaneous radio-link quality is never possible. *Hybrid ARQ*, which requests retransmission of erroneously received data packets, is therefore useful. This can be seen as a mechanism for handling variations in the instantaneous radio-link quality *after* transmission and nicely complements channel-dependent scheduling and link adaptation. Hybrid ARQ also serves the purpose of handling random errors due to, for example, noise in the receiver.

4.5.1 Link adaptation: Power and rate control

Historically, *dynamic transmit-power control* has been used in CDMA-based mobile-communication systems such as WCDMA and cdma2000 to compensate for variations in the instantaneous channel conditions. As the name suggests, dynamic power control dynamically adjusts the radio-link transmit power to compensate for variations and differences in the instantaneous channel conditions. The aim of these adjustments is to maintain a (near) constant E_b/N_0 at the receiver to successfully transmit data without a too high error probability. In principle, transmit-power control increases the power at the transmitter when the radio link experiences poor radio conditions (and vice versa). Thus, the transmit power is in essence inversely proportional to the channel quality as illustrated in Figure 4.5-1a. This results in a basically constant data rate, regardless of the channel variations. For services such as circuit-switched voice, this is a desirable property. Transmit-power control can be seen as one type of link adaptation, that is the adjustment of transmission parameters, in this case the transmit power, to adapt to differences and variations in the instantaneous channel conditions to maintain the received E_b/N_0 at a desired level.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved

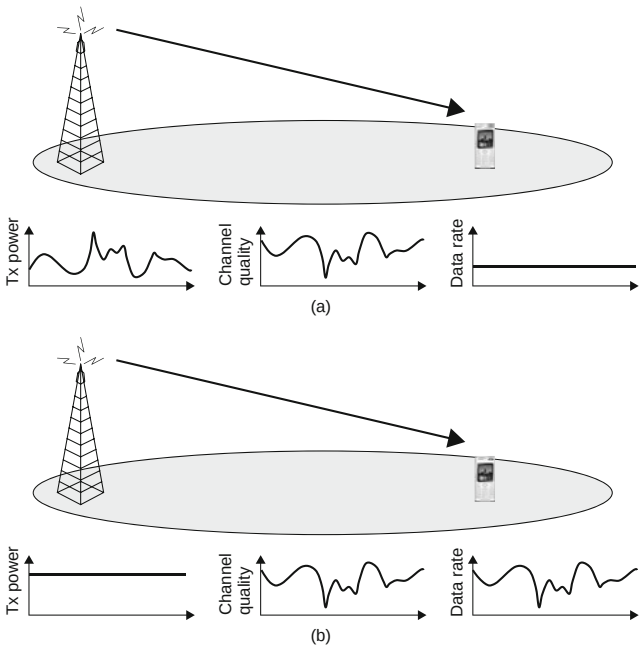


Figure 4.5-1 (a) Power control and (b) rate control.

However, in many cases of mobile communication, especially in case of packet-data traffic, there is not a strong need to provide a certain constant data rate over a radio link. Rather, from a user perspective, the data rate provided over the radio interface should simply be as 'high as possible.' Actually, even in case of typical 'constant-rate' services such as voice and video, (short-term) variations in the data rate are often not an issue, as long as the average data rate remains constant, assuming averaging over some relatively short time interval. In such cases, that is when a constant data rate is not required, an alternative to transmit-power control is link adaptation by means of *dynamic rate control*. Rate control does not aim at keeping the instantaneous radio-link data rate constant, regardless of the instantaneous channel conditions. Instead, with rate control, the data rate is dynamically adjusted to compensate for the varying channel conditions. In situations with advantageous channel conditions, the data rate is increased and vice versa. Thus, rate control maintains the $E_b/N_0 \sim P/R$ at the desired level, not by adjusting the transmission power P, but rather by adjusting the data rate R. This is illustrated in Figure 4.5-1b.

It can be shown that rate control is more efficient than power control. Rate control in principle implies that the power amplifier is always transmitting at full power and therefore efficiently utilized. Power control, on the other hand, results in the power amplifier in most situations not being efficiently utilized as the transmission power is less than its maximum.

In practice, the radio-link data rate is controlled by adjusting the modulation scheme and/or the channel coding rate. In case of advantageous radio-link conditions, the E_b/N_0 at the receiver is high and the main limitation of the data rate is the bandwidth of the radio link. Hence, in such situations higher-order modulation, for example 16QAM or 64QAM, together with a high code rate is appropriate as discussed in Chapter 4.3. Similarly, in case of poor radio-link conditions, QPSK and low-rate coding is used. For this reason, link adaptation by means of rate control is sometimes also referred to as *Adaptive Modulation and Coding* (AMC).

4.5.2 Channel-dependent scheduling

Scheduling controls the allocation of the shared resources among the users at each time instant. It is closely related to link adaptation and often scheduling and link adaptation is seen as one joint function. The scheduling principles, as well as which resources that are shared between users, differ depending on the radio-interface characteristics, for example, whether uplink or downlink is considered and whether different users' transmissions are mutually orthogonal or not.

4.5.2.1 Downlink scheduling

In the downlink, transmissions to different mobile terminals within a cell are typically mutually orthogonal, implying that, at least in theory, there is no interference between the transmissions (no intra-cell interference). Downlink intra-cell orthogonality can be achieved in time domain, *Time Division Multiplexing* (TDM); in the frequency domain, *Frequency-Domain Multiplexing* (FDM); or in the code domain, *Code Domain Multiplexing* (CDM). In addition, the spatial domain can also be used to separate users, at least in a quasi-orthogonal way, through different antenna arrangements. This is sometimes referred to as *Spatial Division Multiplexing* (SDM), although it, in most cases, is used in combination with one or several of the above multiplexing strategies and not discussed further in this chapter.

For packet data, where the traffic often is very bursty, it can be shown that TDM is preferable from a theoretical point-of-view and is therefore typically the main component in the downlink. However, the TDM component is often combined with sharing of the radio resource also in the frequency domain (FDM) or in the code domain (CDM). For example, in case of HSDPA, downlink multiplexing is a combination of TDM and CDM. On the other hand, in case of LTE, downlink multiplexing is a combination of TDM and FDM. The reasons for sharing the resources in domains other than

the time domain will also be elaborated further later in this section.

When transmissions to multiple users occur in parallel, either by using FDM or CDM, there is also an instantaneous sharing of the total available cell transmit power. In other words, not only the time/frequency/code resources are shared resources, but also the power resource in the base station. In contrast, in case of sharing only in the time domain there is, per definition, only a single transmission at a time and thus no instantaneous sharing of the total available cell transmit power.

For the purpose of discussion, assume initially a TDM-based downlink with a single user being scheduled at a time. In this case, the utilization of the radio resources is maximized if, at each time instant, all resources are assigned to the user with the best instantaneous channel condition:

- In case of link adaptation based on power control, this implies that the lowest possible transmit power can be used for a given data rate and thus minimizes the interference to transmissions in other cells for a given link utilization.
- In case of link adaptation based on rate control, this implies that the highest data rate is achieved for a given transmit power, or, in other words, for a given interference to other cells, the highest link utilization is achieved.

However, if applied to the downlink, transmit-power control in combination with TDM scheduling implies that the total available cell transmit power will, in most cases, not be fully utilized. Thus, rate control is generally preferred.

The strategy outlined above is an example of channel-dependent scheduling, where the scheduler takes the instantaneous radio-link conditions into account. Scheduling the user with the instantaneously best radio link conditions is often referred to as *max-C/I* (or *maximum rate*) scheduling. Since the radio conditions for the different radiolink within a cell typically vary independently, at each point in time there is almost always a radio link whose channel quality is near its peak (see Figure 4.5-2). Thus, the channel eventually used for transmission will typically have a high quality and, with rate control, a correspondingly high data rate can be used. This translates into a high system capacity. The gain obtained by transmitting to users with favorable radio-link conditions is commonly known as multi-user diversity; the gains are larger, the larger the channel variations and the larger the number of users in a cell. Hence, in contrast to the traditional view that fast fading, that is, rapid variations in the radio-link quality, is an undesirable effect that has to be combated, with the possibility for channel-dependent scheduling *fading is in fact potentially beneficial and should be exploited.*

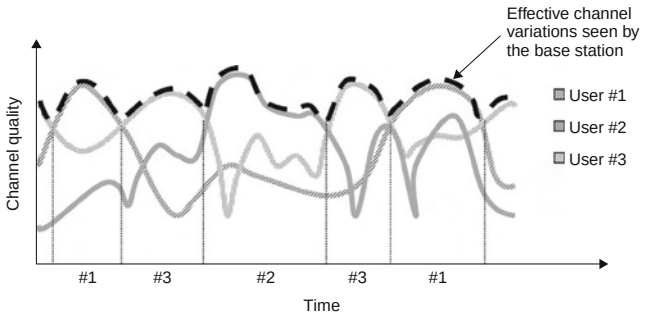


Figure 4.5-2 Channel-dependent scheduling.

Mathematically, the max-C/I (maximum rate) scheduler can be expressed as scheduling user k given by

$$k = \arg \max_i R_i$$

where R_i is the instantaneous data rate for user i. Although, from a system capacity perspective, max-C/I scheduling is beneficial, this scheduling principle will not be fair in all situations. If all mobile terminals are, on average, experiencing similar channel conditions and large variations in the instantaneous channel conditions are only due to, for example, fast multi-path fading, all users will experience the same average data rate. Any variations in the instantaneous data rate are rapid and often not even noticeable by the user. However, in practice different mobile terminals will experience also differences in the (short-term) average channel conditions, for example, due to differences in the distance and shadow fading between the base station and the mobile terminal. In this case, the channel conditions experienced by one mobile terminal may, for a relatively long time, be worse than the channel conditions experienced by other mobile terminals. A pure max-C/I-scheduling strategy may then, in essence, 'starve' the mobile terminal with the bad channel conditions, and the mobile terminal with bad channel conditions will never be scheduled. This is illustrated in Figure 4.5-3a where a max-C/I scheduler is used to schedule between two different users with different average channel quality. Virtually all the time the same user is scheduled. Although resulting in the highest system capacity, this situation is often not acceptable from a quality-of-service point-of-view.

An alternative to the max-C/I scheduling strategy is so-called *round-robin* scheduling, illustrated in Figure 4.5-3b. This scheduling strategy let the users take turns in using the shared resources, without taking the instantaneous channel conditions into account. Round-robin scheduling can be seen as fair scheduling in the sense that the same amount of radio resources (the same amount of time) is given to each communication link.

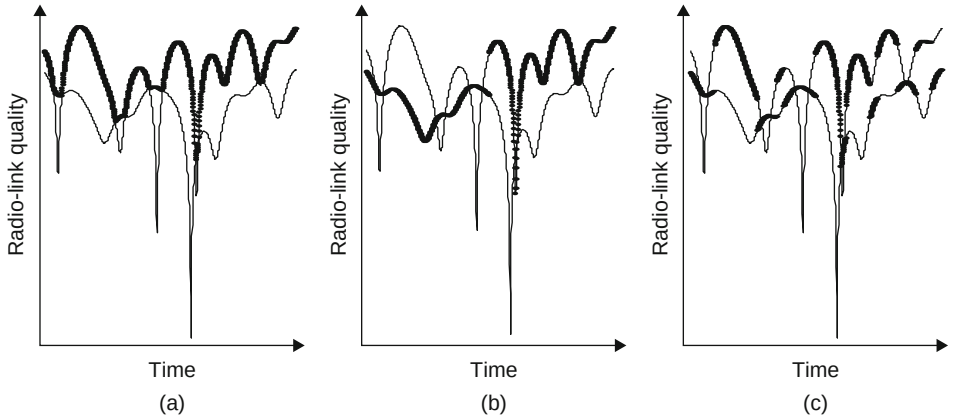


Figure 4.5-3 Example of three different scheduling behaviors for two users with different average channel quality: (a) max C/I, (b) round robin and (c) proportional fair. The selected user is shown with bold lines.

However, round-robin scheduling is not fair in the sense of providing the same service quality to all communication links. In that case more radio resources (more time) must be given to communication links with bad channel conditions. Furthermore, as round-robin scheduling does not take the instantaneous channel conditions into account in the scheduling process it will lead to lower overall system performance but more equal service quality between different communication links, compared to max-C/I scheduling.

Thus what is needed is a scheduling strategy that is able to utilize the fast channel variations to improve the overall cell throughput while still ensuring the same average user throughput for all users or at least a certain minimum user throughput for all users. When discussing and comparing different scheduling algorithms it is important to distinguish between different types of variations in the service quality:

- Fast variations in the service quality corresponding to, for example, fast multipath fading and fast variations in the interference level. For many packet-data applications, relatively large short-term variations in service quality are often acceptable or not even noticeable to the user.
- More long-term differences in the service quality between different communication links corresponding to, for example, differences in the distance to the cell site and shadow fading. In many cases there is a need to limit such long-term differences in service quality.

A practical scheduler should thus operate somewhere in-between the max-C/I scheduler and the round-robin scheduler, that is try to utilize fast variations in channel conditions as much as possible while still satisfying some degree of fairness between users.

One example of such a scheduler is the proportional-fair scheduler, illustrated in Figure 4.5-3c. In this strategy, the shared resources are assigned to the user with the *relatively* best radio-link conditions, that is, at each time instant user k is selected for transmission according to

$$k = \arg \max_i \frac{R_i}{\overline{R}_i}$$

where R_i is the instantaneous data rate for user i and $\overline{R}_i$ is the average data rate for user i. The average is calculated over a certain averaging period T_{PF}. To ensure efficient usage of the short-term channel variations and, at the same time, limit the long-term differences in service quality to an acceptable level, the time constant T_{PF} should be set longer than the time constant for the short-term variations. At the same time T_{PF} should be sufficiently short so that quality variations within the interval T_{PF} are not strongly noticed by a user. Typically, T_{PF} can be set to be in the order of one second.

In the above discussion, it was assumed that all the radio resources in the downlink were assigned to a single user at a time, that is scheduling were done purely in the time domain using TDM between users. However, in several situations, TDM is complemented by CDM or FDM. In principle, there are two reasons for not relying solely on TDM in the downlink:

- In case of insufficient payload, that is the amount of data to transfer to a user is not sufficiently large to utilize the full channel capacity, and a fraction of resources could be assigned to another user, either through FDM or CDM.
- In case channel variations in the frequency domain are exploited through FDM as discussed further below.

The scheduling strategies in these cases can be seen as generalizations of the schemes discussed for the TDM-only cases above. For example, to handle small

payloads, a greedy filling approach can be used, where the scheduled user is selected according to max-C/I (or any other scheduling scheme). Once this user has been assigned resources matching the amount of data awaiting transmission, the second best user according to the scheduling strategy is selected and assigned (a fraction of) the residual resources and so on.

Finally, it should also be noted that the scheduling algorithm typically is a base-station-implementation issue and nothing that is normally specified in any standard. What needs to be specified in a standard to support channel-dependent scheduling is channel-quality measurements/reports and the signaling needed for dynamic resource allocation.

4.5.2.2 Uplink scheduling

The previous section discussed scheduling from a downlink perspective. However, scheduling is equally applicable to uplink transmissions and to a large extent the same principles can be reused although there are some differences between the two.

Fundamentally, the uplink power resource is *distributed* among the users, while in the downlink the power resource is *centralized* within the base station. Furthermore, the maximum uplink transmission power of a single terminal is typically significantly lower than the output power of a base station. This has a significant impact on the scheduling strategy. Unlike the downlink, where pure TDMA often can be used, uplink scheduling typically has to rely on sharing in the frequency and/or code domain in addition to the time domain as a single terminal may not have sufficient power for efficiently utilizing the link capacity.

Channel-dependent scheduling is, similar to the downlink case, beneficial also in the uplink case. However, the characteristics of the underlying radio interface, most notably whether the uplink relies on orthogonal or non-orthogonal multiple access and the type of link adaptation scheme used, also have a significant impact on the uplink scheduling strategy.

In case of a non-orthogonal multiple-access scheme such as CDMA, power control is typically essential for proper operation. As discussed earlier in this chapter, the purpose of power control is to control the received E_b/N_0 such that the received information can be recovered. However, in a non-orthogonal multiple-access setting, power control also serves the purpose of controlling the amount of interference affecting *other* users. This can be expressed as the *maximum tolerable interference level* at the base station is a shared resource. Even if it, from a single user's perspective, would be beneficial to transmit at full power to maximize the data rate, this may not be acceptable from an interference perspective as other terminals in this case may not be able to successfully transfer any data. Thus, with non-orthogonal multiple access, scheduling a terminal when the channel conditions are favorable may not directly translate into a higher data rate as the interference generated to other simultaneously transmitting terminals in the cell must be taken into account. Stated differently, the *received* power (and thus the data rate) is, thanks to power control, in principle constant, regardless of the channel conditions at the time of transmission, while the *transmitted* power depends on the channel conditions at the time of transmission. Hence, even though channel-dependent scheduling in this example does not give a direct gain in terms of a higher data rate from the terminal, channel-dependent scheduling will still provide a gain for the system in terms of lower intra-cell interference.

The above discussion on non-orthogonal multiple access was simplified in the sense that no bounds on the terminals transmission power were assumed. In practice, the transmission power of a terminal is upper-bounded, both due to implementation and regulatory reasons, and scheduling a terminal for transmission in favorable channel conditions decreases the probability that the terminal has insufficient power to utilize the channel capacity.

In case of orthogonal multiple-access scheme, intra-cell power control is fundamentally not necessary and the benefits with channel-dependent scheduling become more similar to the downlink case. In principle, from an intra-cell perspective, a terminal can transmit at full power and the scheduler assigns a suitable part of the orthogonal resources (in practice a suitable part of the overall bandwidth) to the terminal for transmission. The remaining orthogonal resources can be assigned to other users. However, implementation constraints, for example leakage between the received signals or limited dynamic range in the receiver circuitry, may pose restrictions on the maximum tolerable power difference between the signals from simultaneously transmitting terminals. As a consequence, a certain degree of power control may be necessary, making the situation somewhat similar to the non-orthogonal case.

The discussion on non-orthogonal and orthogonal multiple access mainly considered intra-cell multiple access. However, in many practical systems universal frequency reuse between cells is used. In this case, the inter-cell multiple access is non-orthogonal, regardless of the intra-cell multiple access, which sets limits on the allowable transmission power from a terminal.

Regardless of whether orthogonal or non-orthogonal multiple access is used, the same basic scheduling principles as for the downlink can be used. A max-C/I scheduler would assign all the uplink resources to the terminal with the best uplink channel conditions. Neglecting any power limitations in the terminal, this would result in the highest capacity (in an isolated cell).

In case of a non-orthogonal multiple-access scheme, *greedy filling* is one possible scheduling strategy. With greedy filling, the terminal with the best radio conditions is assigned as high data rate as possible. If the interference level at the receiver is smaller than the maximum tolerable level, the terminal with the second best channel conditions is allowed to transmit as well, continuing with more and more terminals until the maximum tolerable interference level at the receiver is reached. This strategy maximizes the air interface utilization but is achieved at the cost of potentially large differences in data rates between users. In the extreme case, a user at the cell border with poor channel conditions may not be allowed to transmit at all.

Strategies between greedy filling and max-C/I can also be envisioned, for example different proportional-fair strategies. This can be achieved by including a weighting factor for each user, proportional to the ratio between the instantaneous and average data rates, into the greedy filling algorithm.

The schedulers above all assume knowledge of the instantaneous radio-link conditions; knowledge that can be hard to obtain in the uplink scenario as discussed in Section 4.5.2.4 below. In situations when no information about the uplink radio-link quality is available at the scheduler, *round-robin* scheduling can be used. Similar to the downlink, round-robin implies terminals taking turns in transmitting, thus creating a TDMA-like operation with inter-user orthogonality in the time domain. Although the round-robin scheduler is simple, it is far from the optimal scheduling strategy.

However, the transmission power in a terminal is limited and therefore additional sharing of the uplink resources in frequency and/or code domain is required. This also impacts the scheduling decisions. For example, terminals far from the base station typically operate in the power-limited region, in contrast to terminals close to the base stations which often are in the bandwidth-limited region (for a discussion on power-limited vs. bandwidth-limited operation, see Chapter 4.3). Thus, for a terminal far from the base station, increasing the bandwidth will not result in an increased data rate and it is better to only assign a small amount of the bandwidth to this terminal and assign the remaining bandwidth to other terminals. On the other hand, for terminals close to the base station, an increase in the assigned bandwidth will provide a higher data rate.

4.5.2.3 Link adaptation and channel-dependent scheduling in the frequency domain

In the previous section, TDM-based scheduling was assumed and it was explained how, in this case, channel variations in the time domain could be utilized to improve system performance by applying channel-dependent scheduling, especially in combination with dynamic rate control. However, if the scheduler has access to the frequency domain, for example through the use of OFDM transmission, scheduling and link adaptation can also take place in the frequency domain.

Link adaptation in the frequency domain implies that, based on knowledge about the instantaneous channel conditions in the frequency domain, that is knowledge about the attenuation as well as the noise/interference level of, in the extreme case, every OFDM subcarrier, the power and/or the data rate of each OFDM carrier can be individually adjusted for optimal utilization.

Similarly, channel-dependent scheduling in the frequency domain implies that, based on knowledge about the instantaneous channel conditions in the frequency domain, different subcarriers are used for transmission to or from different mobile terminals. The scheduling gains from exploiting variations in the frequency-domain are similar to those obtained from time-domain variations. Obviously, in situations where the channel quality varies significantly with the frequency while the channel quality only varies slowly with time, channel-dependent scheduling in the frequency domain can enhance system capacity. An example of such a situation is a wideband indoor system with low mobility, where the quality only varies slowly with time.

4.5.2.4 Acquiring on channel-state information

To select a suitable data rate, in practice a suitable modulation scheme and channel-coding rate, the transmitter needs information about the radio-link channel conditions. Such information is also required for the purpose of channel-dependent scheduling. In case of a system based on frequency-division duplex (FDD), only the receiver can accurately estimate the radio-link channel conditions.

For the downlink, most systems provide a downlink signal of a predetermined structure, known as the downlink pilot or the downlink reference signal. This reference signal is transmitted from the base station with a constant power and can be used by the mobile terminal to estimate the instantaneous downlink channel conditions. Information about the instantaneous downlink conditions can then be reported to the base station.

Basically, what is relevant for the transmitter is an estimate reflecting the channel conditions *at the time of transmission*. Hence, in principle, the terminal could apply a predictor, trying to predict the future channel conditions and report this predicted value to the base

station. However, as this would require specification of prediction algorithms and how they would operate when the terminal is moving at different speeds, most practical systems simply report the measured channel conditions to the base station. This can be seen as a very simple predictor, basically assuming the conditions in the near future will be similar to the current conditions. Thus, the more rapid the time-domain channel variations are, the less efficient link adaptation is.

As there inevitably will be a delay between the point in time when the terminal measured the channel conditions and the application of the reported value in the transmitter, channel-dependent scheduling and link adaptation typically operates at its best at low terminal mobility. If the terminal starts to move at a high speed, the measurement reports will be outdated once reported to the base station. In such cases, it is often preferable to perform link adaptation on the long-term average channel quality and rely on hybrid ARQ with soft combining for the rapid adaptation.

For the uplink, estimation of the uplink channel conditions is not as straightforward as there is typically not any reference signal transmitted with constant power from each mobile terminal.

In case of a system with time-division duplex (TDD) where uplink and downlink communication are time multiplexed within the same frequency band, the uplink signal attenuation could be estimated from downlink measurements of the mobile terminal, due to the reciprocity of the multi-path fading in case of TDD. However, it should then be noted that this may not provide full knowledge of the downlink channel conditions. As an example, the interference situations at the mobile terminal and the base station are different also in case of TDD.

4.5.2.5 Traffic behavior and scheduling

It should be noted that there is little difference between different scheduling algorithms at low system load, that is when only one or, in some cases, a few users have data waiting for transmission at the base station at each scheduling instant. The differences between different scheduling algorithms are primarily visible at high load. However, not only the load, but also the traffic behavior affects the overall scheduling performance.

As discussed above, channel-dependent scheduling tries to exploit short-term variations in radio quality. Generally speaking, a certain degree of long-term fairness in service quality is desirable, which should be accounted for in the scheduler design. However, since system throughput decreases the more fairness is enforced, a trade-off between fairness and system throughput is necessary. In this trade-off, it is important to take traffic characteristics into account as they have a significant influence on the trade-off between system throughput and service quality.

To illustrate this, consider three different downlink schedulers:

1. *Round-robin (RR) scheduler,* where channel conditions are not taken into account.
2. *Proportional-fair (PF) scheduler,* where short-term channel variations are exploited while maintaining the long-term average user data rate.
3. *Max-C/I scheduler,* where the user with the best instantaneous channel quality in absolute terms is scheduled.

For a full buffer scenario when there is always data available at the base station for all terminals in the cell, a max-C/I scheduler will result in no, or a very low, user throughput for users at the cell edge with a low average channel quality. The reason is the fundamental strategy of the max-C/I scheduler – all resources are allocated for transmission to the terminal whose channel conditions support the highest data rate. Only in the rare, not to say unlikely, case of a cell-edge user having better conditions than a cell-center user, for example due to a deep fading dip for the cell-center user, will the cell-edge user be scheduled. A proportional-fair scheduler, on the other hand, will ensure some degree of fairness by selecting the user supporting the highest data rate relative to its average data rate. Hence, users tend to be scheduled on their fading peaks, regardless of the absolute quality. Thus, users on the cell edge will also be scheduled, thereby resulting in (some degree of) fairness between users.

For a scenario with bursty packet data, the situation is different. In this case, the users' buffers will be finite and in many cases also empty. For example, a web page has a certain size and after transmitting the page, there is no more data to be transmitted to the terminal in question until the users requests a new page by clicking on a link. In this case, a max-C/I scheduler can still provide a certain degree of fairness. Once the buffer for the user with the highest C/I has been emptied, another user with non-empty buffers will have the highest C/I and be scheduled and so on. This is the reason for the difference between full buffer and web-browsing traffic illustrated in Figure 4.5-4. The proportional-fair scheduler has similar performance in both scenarios.

Clearly, the degree of fairness introduced by the traffic properties depends heavily on the actual traffic; a design made with certain assumptions may be less desirable in an actual network where the traffic pattern may be different from the assumptions made during the design. Therefore, relying solely on the traffic properties for fairness is not a good strategy, but the discussion above

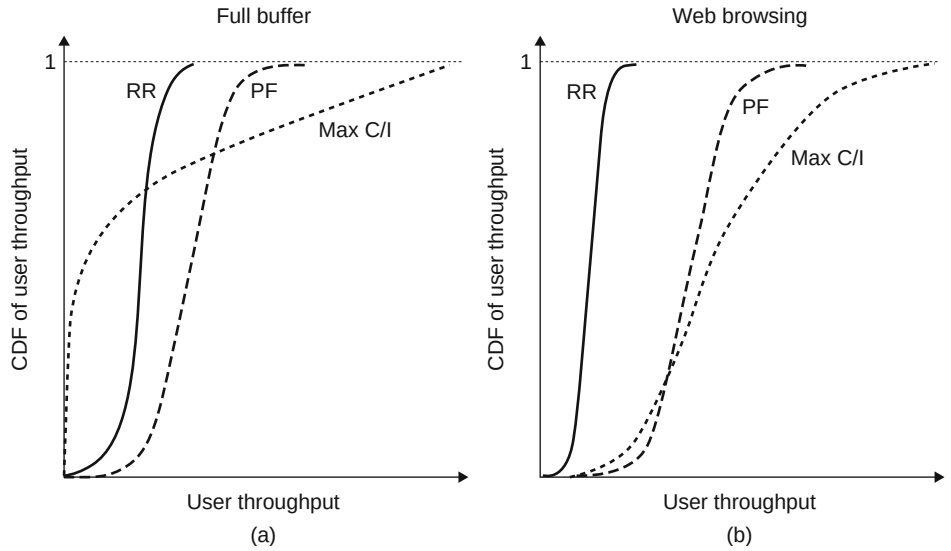


Figure 4.5-4 Illustration of the principle behavior of different scheduling strategies: (a) for full buffers and (b) for web browsing traffic model.

emphasizes the need to not only design the scheduler for the full buffer case.

4.5.3 Advanced retransmission schemes

Transmissions over wireless channels are subject to errors, for example due to variations in the received signal quality. To some degree, such variations can be counteracted through link adaptation as discussed above. However, receiver noise and unpredictable interference variations cannot be counteracted. Therefore, virtually all wireless communications systems employ some form of *Forward Error Correction* (FEC), tracing its roots to the pioneering work by Claude Shannon in 1948. There is a rich literature in the area of error-correcting coding, and a detailed description is beyond the scope of this book. In short, the basic principle beyond forward error-correcting coding is to introduce redundancy in the transmitted signal. This is achieved by adding *parity bits* to the information bits prior to transmission (alternatively, the transmission could consists of parity bits alone, depending on the coding scheme used). The parity bits are computed from the information bits using a method given by the coding structure used. Thus, the number of bits transmitted over the channel is larger then the number of original information bits and a certain amount of *redundancy* has been introduced in the transmitted signal.

Another approach to handle transmissions errors is to use *Automatic Repeat Request* (ARQ). In an ARQ scheme, the receiver uses an error-detecting code, typically a Cyclic Redundancy Check (CRC), to detect if the received packet is in error or not. If no error is detected in the received data packet, the received data is declared error-free and the transmitter is notified by sending a positive acknowledgment (ACK). On the other hand, if an error is detected, the receiver discards the received data and notifies the transmitter via a return channel by sending a negative acknowledgment (NAK). In response to a NAK, the transmitter retransmits the same information.

Virtually all modern communication systems, including WCDMA and cdma2000, employ a combination of forward error-correcting coding and ARQ, a combination known as *hybrid ARQ*. Hybrid ARQ uses forward error correcting codes to correct a subset of all errors and rely on error detection to detect uncorrectable errors. Erroneously received packets are discarded and the receiver requests retransmissions of corrupted packets. Thus, it is a combination of FEC and ARQ as described above. Most practical hybrid ARQ schemes are built around a CRC code for error detection and convolutional or Turbo codes for error correction, but in principle any error-detecting and error-correcting code can be used.

4.5.4 Hybrid ARQ with soft combining

The hybrid ARQ operation described above discards erroneously received packets and requests retransmission. However, despite the fact that the packet was not possible to decode, the received signal contains information, which is lost by discarding erroneously received packets. This shortcoming is addressed by *hybrid ARQ with soft*

combining. In hybrid ARQ with soft combining, the erroneously received packet is stored in a buffer memory and later combined with the retransmission to obtain a single, combined packet which is more reliable than its constituents. Decoding of the error-correcting code operates on the combined signal. If the decoding fails (typically a CRC code is used to detect this event), a retransmission is requested.

Retransmission in any hybrid ARQ scheme must, by definition, represent the same set of information bits as the original transmission. However, the set of coded bits transmitted in each retransmission may be selected differently as long as they represent the same set of information bits. Hybrid ARQ with soft combining is therefore usually categorized into C*hase combining* and *Incremental Redundancy*, depending on whether the retransmitted bits are required to be identical to the original transmission or not.

In C*hase combining*, the retransmissions consist of the same set of coded bits as the original transmission. After each retransmission, the receiver uses maximum-ratio combining to combine each received channel bit with any previous transmissions of the same bit and the combined signal is fed to the decoder. As each retransmission is an identical copy of the original transmission, retransmissions with Chase combining can be seen as additional repetition coding. Therefore, as no new redundancy is transmitted, Chase combining does not give any additional coding gain but only increases the accumulated received E_b/N_0 for each retransmission (Figure 4.5-5).

Several variants of Chase combining exist. For example, only a subset of the bits transmitted in the original transmission might be retransmitted, so-called partial Chase combining. Furthermore, although combining is often done after demodulation but before channel decoding, combining can also be done at the modulation symbol level before demodulation, as long as the modulation scheme is unchanged between transmission and retransmission.

With *Incremental Redundancy* (IR), each retransmission does not have to be identical to the original transmission. Instead, *multiple sets* of coded bits are generated, each of them representing the same set of information bits. Whenever a retransmission is required, the retransmission typically uses a different set of coded bits than the previous transmission. The receiver combines the retransmission with previous transmission attempts of the same packet. As the retransmission may contain additional parity bits, not included in the previous transmission attempts, the resulting code rate is generally lowered by a retransmission. Furthermore, each retransmission does not necessarily have to consist of the same number of coded bits as the original and, in general, also the modulation scheme can be different for different retransmissions. Hence, incremental redundancy can be seen as a generalization of Chase combining or, stated differently, Chase combining is a special case of incremental redundancy.

Typically, incremental redundancy is based on a low-rate code and the different redundancy versions are generated by puncturing the output of the encoder. In the first transmission only a limited number of the coded bits are transmitted, effectively leading to a high-rate code. In the retransmissions, additional coded bits are transmitted. As an example, assume a basic rate-1/4 code. In the first transmission, only every third coded bit is transmitted, effectively giving a rate-3/4 code as illustrated in Figure 4.5-6. In case of a decoding error and a subsequent request for a retransmission, additional bits are transmitted, effectively leading to a rate-3/8 code. After a second retransmission the code rate is 1/4. In case of more than two retransmissions, already transmitted coded bits would be repeated. In addition to a gain in accumulated received E_b/N_0, incremental redundancy

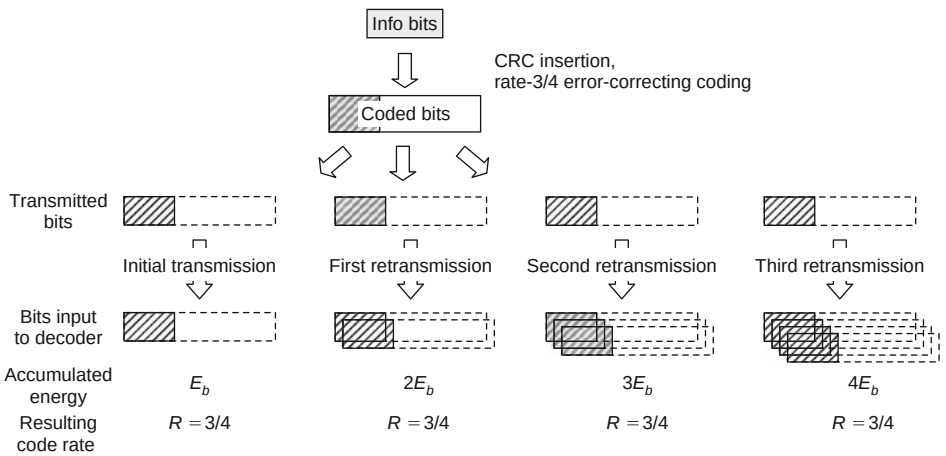


Figure 4.5-5 Example of chase combining.

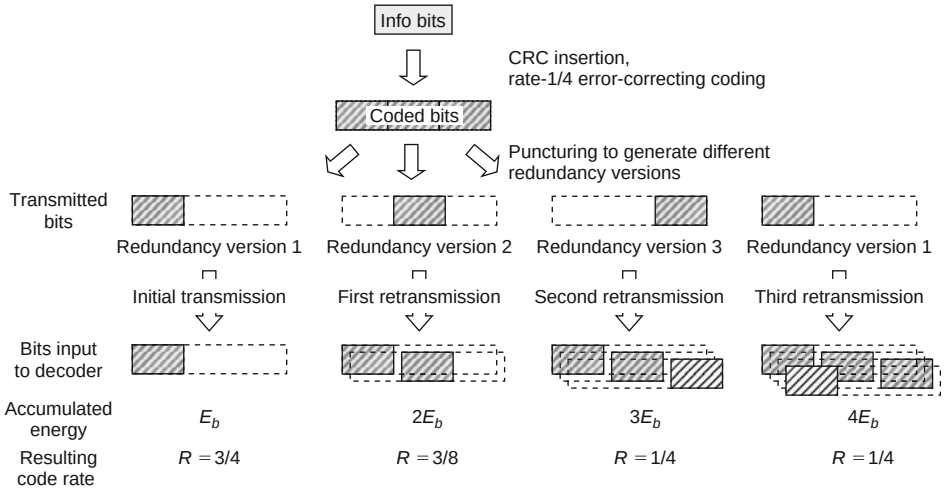


Figure 4.5-6 Example of incremental redundancy.

also results in a coding gain for each retransmission. The gain with IR compared to Chase is larger for high initial code rates while at lower initial coding rates, Chase combining is almost as good as IR. Furthermore, the performance gain of incremental redundancy compared to Chase combining can also depend on the relative power difference between the transmission attempts.

With incremental redundancy, the code used for the first transmission should provide good performance not only when used alone, but also when used in combination with the code for the second transmission. The same holds for subsequent retransmissions. Thus, as the different redundancy versions typically are generated through puncturing of a low-rate mother code, the puncturing patterns should be defined such that all the code bits used by a high-rate code should also be part of any lower-rate codes. In other words, the resulting code rate R_i after transmission attempt i, consisting of the coded bits from redundancy versions RV_k, $k = 1,\ldots, i$, should have similar performance as a good code designed directly for rate R_i. Examples of this for convolutional codes are the so-called rate-compatible convolutional codes.

In the discussion so far, it has been assumed that the receiver has received all the previously transmitted redundancy versions. If all redundancy versions provide the same amount of information about the data packet, the order of the redundancy versions is not critical. However, for some code structures, not all redundancy versions are of equal importance. One example hereof is Turbo codes, where the systematic bits are of higher importance than the parity bits. Hence, the initial transmission should at least include all the systematic bits and some parity bits. In the retransmission(s), parity bits not part of the initial transmission can be included. However, if the initial transmission was received with poor quality or not at all, a retransmission with only parity bits is not appropriate as a retransmission of (some of) the systematic bits provides better performance. Incremental redundancy with Turbo codes can therefore benefit from multiple levels of feedback, for example by using two different negative acknowledgments - NAK to request additional parity bits and LOST to request a retransmission of the systematic bits. In general, the problem of determining the amount of systematic and parity bits in a retransmission based on the signal quality of previous transmission attempts is non-trivial.

Hybrid ARQ with soft combining, regardless of whether Chase or incremental redundancy is used, leads to an implicit reduction of the data rate by means of retransmissions and can thus be seen as *implicit* link adaptation. However, in contrast to link adaptation based on explicit estimates of the instantaneous channel conditions, hybrid ARQ with soft combining implicitly adjusts the coding rate based on the result of the decoding. In terms of overall throughput this kind of implicit link adaptation can be superior to explicit link adaptation as additional redundancy is only added *when needed*, that is when previous higher-rate transmissions were not possible to decode correctly. Furthermore, as it does not try to predict any channel variations, it works equally well, regardless of the speed of which the terminal is moving. Since implicit link adaptation can provide a gain in system throughput, a valid question is why explicit link adaptation at all is necessary. One major reason for having explicit link adaptation is the reduced delay. Although relying on implicit link adaptation alone is sufficient from a system throughput perspective, the end-user service quality may not be acceptable from a delay perspective.

Chapter 4.6

4.6

WCDMA evolution: HSPA and MBMS

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

The first step in the evolution of WCDMA radio access is the introduction of *High-Speed Downlink Packet Access* (HSDPA) in Release 5 of the 3GPP/WCDMA specifications. Although packet-data communication is supported already in the first release of the WCDMA standard, HSDPA brings further enhancements to the provisioning of packet-data services in WCDMA, both in terms of system and end-user performance. The downlink packet-data enhancements of HSDPA are complemented by *Enhanced Uplink*, introduced in Release 6 of the 3GPP/WCDMA specifications. HSDPA and Enhanced Uplink are often jointly referred to as *High-Speed Packet Access* (HSPA).

As discussed in Chapter 4.2, important requirements for cellular systems providing packet-data services are high data rates and low delays while, as at the same time, maintaining good coverage and providing high capacity. To achieve this, HSPA introduces several of the basic techniques into WCDMA, such as higher order modulation, fast (channel-dependent) scheduling and rate control, and fast hybrid ARQ with soft combining. Altogether, HSPA provides downlink and uplink data rates up to approximately 14 and 5.7 Mbit/s, respectively, and significantly reduced round trip times and improved capacity, compared to Release 99.

3GPP Release 6 also brings efficient support for broadcast services into WCDMA through the introduction of *Multimedia Broadcast Multicast Services* (MBMS), suitable for applications like mobile TV. With MBMS, multiple terminals may receive the same broadcast transmission instead of the network transmitting the same information individually to each of the users. Naturally, this will lead to an improvement in the resource utilization when several users in a cell receive the same content. As a consequence of the broadcast transmission, user-specific adaptation of the transmission parameters cannot be used and diversity as a mean to maintain good coverage is important. For MBMS, macro-diversity through multi-cell reception is employed for this purpose.

The evolution of the WCDMA radio access continues and will continue also in the future. For example, 3GPP Release 7 introduces several new features. MIMO is a tool to further improve capacity and especially the HSPA peak data rates. *Continuous Packet Connectivity* aims at providing an 'always-on' service perception terminals.

In parallel to enhancements of the radio-access specifications, there is an ongoing effort in *advanced* receiver structures. These improvements, which can provide a significant gain in both system and end-user performance, are to a large extent implementation specific, although the receiver-performance requirements as such are subject to standardization.

As a summary, the evolution of WCDMA is illustrated in Figure 4.6-1.

The remainder of this chapter contains an overview of the first release of WCDMA. The intention is to provide the reader with an understanding of the foundation upon which HSPA and MBMS are built. The reader already familiar with WCDMA may want to skip the remainder of this chapter.

4.6.1 WCDMA: brief overview

This section provides a brief overview of WCDMA Release 99 to serve as a background to the subsequent sections. As a thorough walk through of WCDMA is a topic of its own and beyond the scope of this book, the reader is referred to other books and articles for a detailed description. WCDMA is a versatile and highly

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved

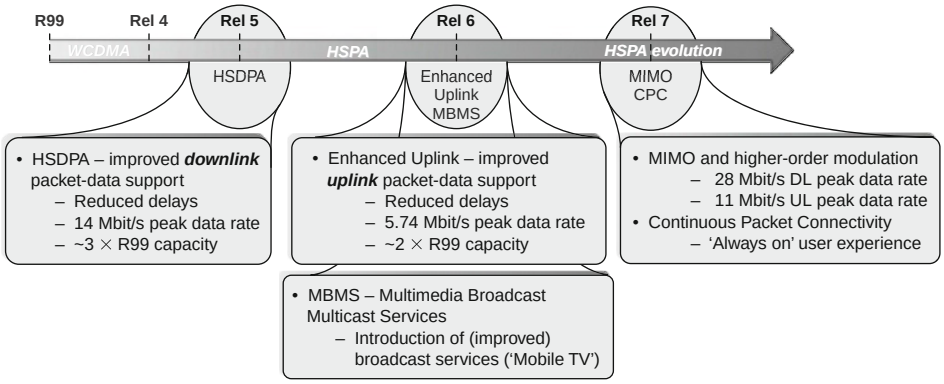


Figure 4.6-1 WCDMA evolution.

flexible radio interface that can be configured to meet the requirements from a large number of services, but the focus for the description below is the functionality commonly used to support packet-data transmissions. The main purpose is to provide a brief background to WCDMA to put the enhancements described later on into a perspective.

4.6.1.1 Overall architecture

WCDMA is based on a hierarchical architecture with the different nodes and interfaces as illustrated in Figure 4.6-2. A terminal, also referred to as *User Equipment* (UE) in 3GPP terminology, communicates with one (or several) NodeBs. In the WCDMA architecture, the term *NodeB* refers to a logical node, responsible for physical-layer processing such as error-correcting coding, modulation and spreading, as well as conversion from baseband to the radio-frequency signal transmitted from the antenna. A NodeB is handling transmission and reception in one or several cells. Three-sector sites are common, where each NodeB is handling transmissions in three cells, although other arrangements of the cells belonging to one NodeB can be thought of, for example, a large number of indoor cell or several cells along a highway belonging to the same NodeB. Thus, a base station is a possible implementation of a NodeB.

The *Radio Network Controller* (RNC) controls multiple NodeBs. The number of NodeBs connected to one RNC varies depending on the implementation and deployment, but up to a few hundred NodeBs per RNC is not uncommon. The RNC is, among other things, in charge of call setup, quality-of-service handling, and management of the radio resources in the cells for which it is responsible. The ARQ protocol, handling retransmissions of erroneous data, is also located in the RNC. Thus, in Release 99, most of the 'intelligence' in the radio-access network resides in the RNC, while the NodeBs mainly act as modems.

Finally, the RNCs are connected to the Internet and the traditional wired telephony network through the core network.

Most modern communication systems structure the processing into different *layers* and WCDMA is no exception. The layered approach is beneficial as it provides a certain structure to the overall processing where each layer is responsible for a specific part of the radio-access functionality. The different protocol layers used in WCDMA are illustrated in Figure 4.6-3 and briefly described.

User data from the core network, for example in the form of IP packets, are first processed by the *Packet Data Convergence Protocol* (PDCP) which performs (optional) header compression. IP packets have a relatively large header, 40 bytes for IPv4 and 60 bytes for IPv6, and to save radio-interface resources, header compression is beneficial.

The *Radio Link Protocol* (RLC) is, among other things, responsible for segmentation of the IP packets into smaller units known as RLC *Protocol Data Units* (RLC PDUs). At the receiving end, the RLC performs the corresponding reassembly of the received segments. The RLC also handles the ARQ protocol. For packet-data services, error-free delivery of data is often essential and the RLC can therefore be configured to request retransmissions of erroneous RLC PDUs in this case. For each incorrectly received PDU, the RLC requests a retransmission. The need for a retransmission is indicated by the RLC entity at the receiving end to its peer RLC entity at the transmitting end by means of status reports.

The *Medium Access Control* (MAC) layer offers services to the RLC in the form of so-called *Logical Channels*. The MAC layer can multiplex data from multiple logical channels. It is also responsible for determining the *Transport Format* of the data sent to the next layer, the *physical layer*. In essence, the transport format is the instantaneous data rate used over the radio link and is better understood once the interface between the MAC and physical layer has been described. This interface is

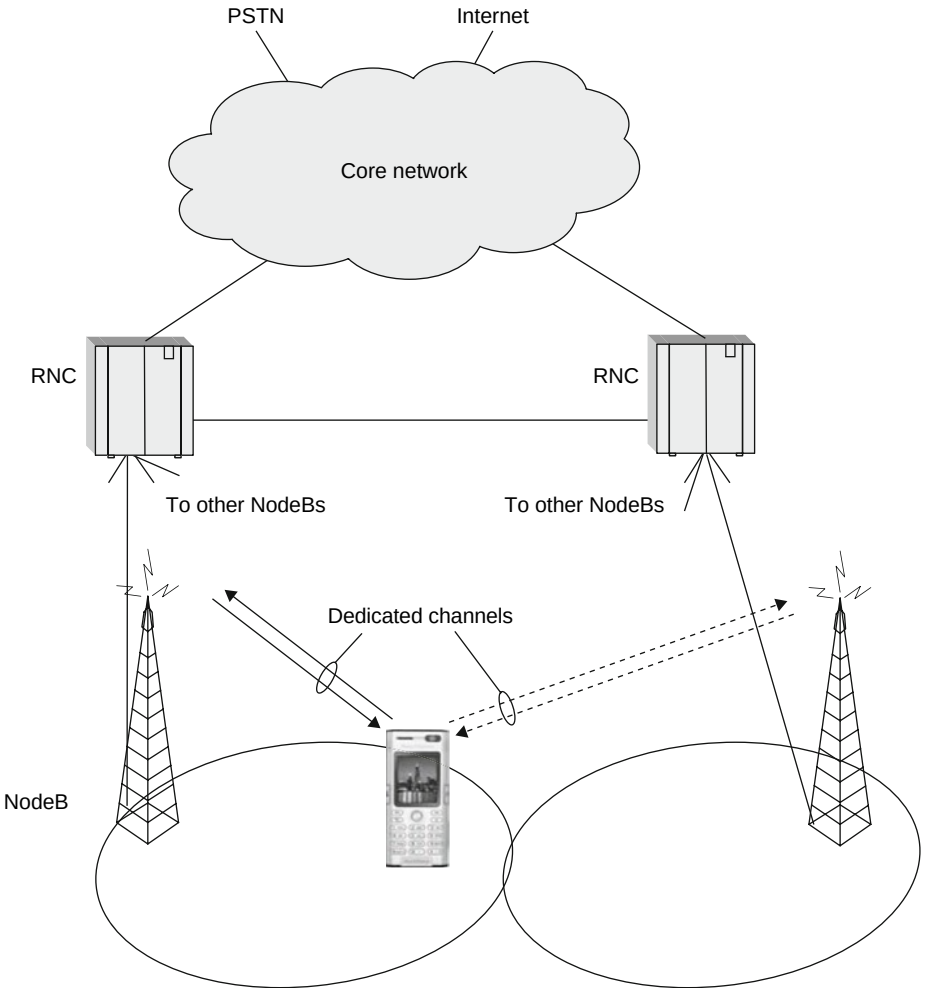


Figure 4.6-2 WCDMA radio-access network architecture.

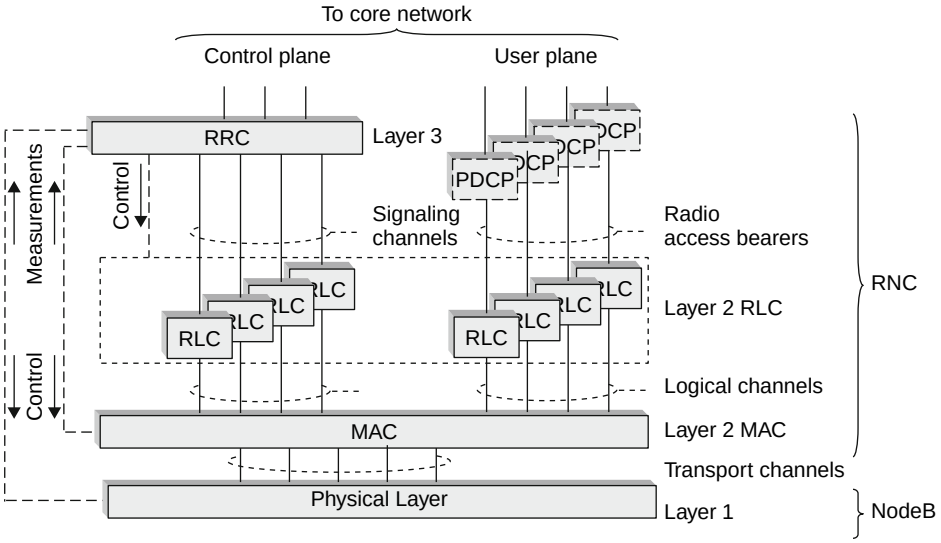


Figure 4.6-3 WCDMA protocol architecture.

specified through so-called *Transport Channels* over which data in the form of *transport blocks* are transferred. There are several transport channels defined in WCDMA.

In each *Transmission Time Interval* (TTI) one or several transport blocks are fed from the MAC layer to the physical layer, which performs coding, interleaving, multiplexing, spreading, etc., prior to data transmission. Thus, for WCDMA, the TTI is the time which the interleaver spans and the time it takes to transmit the transport block over the radio interface. A larger TTI implies better time diversity, but also a longer delay. In the first release, WCDMA relies on TTI lengths of 10, 20, 40, and 80 ms. As will be seen later, HSPA introduces additional 2 ms TTI to reduce the latency.

To support different data rates, the MAC can vary the transport format between consecutive TTIs. The transport format consists of several parameters describing how the data shall be transmitted in a TTI. By varying the transport-block size and/or the number of transport blocks, different data rates can be realized.

At the bottom of the protocol stack is the *physical layer*. The physical layer is responsible for operations such as coding, spreading and data modulation, as well as modulation of the radio-frequency carrier.

PDCP, RLC, MAC, and physical layer, are configured by the *Radio Resource Control* (RRC) protocol. RRC performs admission control, handover decisions, and active set management for soft handover. By setting the parameters of the RLC, MAC, and physical layers properly, the RRC can provide the necessary quality-of-service requested by the core network for a certain service.

On the network side, the MAC, RLC and RRC entities in Release 99 are all located in the RNC while the physical layer is mainly located in the NodeB. The same entities also exist in the UE. For example, the MAC in the UE is responsible for selecting the transport format for uplink transmissions from a set of formats configured by the network. However, the handling of the radio resources in the cell is controlled by the RRC entity in the network and the UE obeys the RRC decisions taken in the network.

4.6.1.2 Physical layer

The basis for the physical layer of WCDMA is spreading of the data to be transmitted to the chip rate, which equals 3.84 Mchip/s. It is also responsible for coding, transport-channel multiplexing, and modulation of the radio-frequency carrier. A simplified illustration of the physical layer processing is provided in Figure 4.6-4.

To each transport block to be transmitted, the transmitter physical layer adds a Cyclic Redundancy Check (CRC) for the purpose of error detection. If the receiver detects an error, the RLC protocol in the receiver is informed and requests a retransmission. After CRC attachment, the data are encoded using a rate-1/3 Turbo coder (convolutional coding is also supported by WCDMA, but typically not used for packet-data services). Rate matching by means of puncturing or repetition of the coded bits is used to fine-tune the code rate and multiple coded and interleaved transport channels can be multiplexed together, forming a single stream of bits to be spread to the chip rate and subsequently modulated. For the downlink, QPSK modulation is used, while the uplink uses BPSK. The resulting stream of modulation symbols is mapped to a *physical channel* and subsequently digital-to-analog converted and modulated onto a radio-frequency carrier. In principle, each physical channel corresponds to a unique spreading code, used to separate transmissions to/from different users.

The spreading operation actually consists of two steps: spreading to the chip rate by using orthogonal *channelization codes* with a length equal to the symbol time followed by scrambling using non-orthogonal scrambling sequences with a length equal to the 10 ms radio frame.

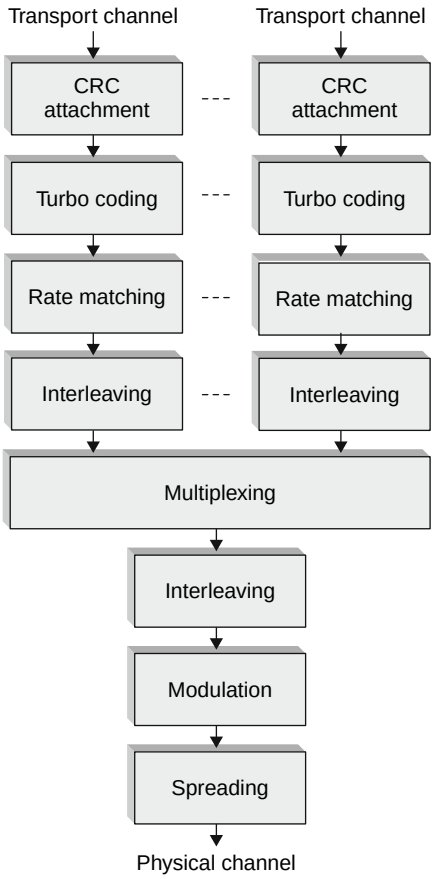


Figure 4.6-4 Simplified view of physical layer processing in WCDMA.

The channelization codes are so-called *Orthogonal Variable Spreading Factor* (OVSF) codes, defined by the tree structure in Figure 4.6-5. A key property of the OVSF codes is the mutual orthogonality between data streams spread with different OVSF codes, even if the data rate, and therefore the spreading factors, are different. This holds as long as the different OVSF codes are selected from different branches of the channelization code tree and transmitted with the same timing. By selecting different spreading factors, physical channels with different data rates can be provided.

In the downlink, data to a certain user, including the necessary control information, are carried on one QPSK-modulated *Dedicated Physical Channel* (DPCH) corresponding to one OVSF code. By varying the spreading factor, different DPCH symbol rates can be provided. The spreading factor is determined by the transport format selected by the MAC layer. There is also a possibility to use multiple physical channels to a single user for the highest data rates.

Some downlink channelization codes are pre-allocated for specific purposes. One of the most important is the code used for reference signal transmission, in WCDMA known as the *Common Pilot Channel* (CPICH). The CPICH contains known data and is used as a reference for downlink channel estimation by all UEs in the cell. There is also a pre-allocated control channel carrying cell-specific control information.

Since physical channels are separated by OVSF codes, transmissions on different physical channels are orthogonal and will not interfere with each other. The WCDMA downlink is therefore often said to be *orthogonal*. However, at the receiver side, orthogonality will partly be lost in case of a frequency-selective channel. This will cause signal corruption to the received signal and result in interference between different codes used for downlink transmission. Such signal corruption can be counteracted by an equalizer.

WCDMA supports asynchronous operation, where transmissions from different cells are not time synchronized. To separate different cells, the scrambling is cell specific in the downlink. A terminal receiving transmissions from one cell will therefore be interfered by transmissions in neighboring cells as the scrambling sequences are non-orthogonal. This interference will be suppressed by the terminal receiver with a factor proportional to the processing gain (the processing gain is given by the spreading factor divided by the code rate after rate matching).

In the uplink, data are carried on a BPSK-modulated *Dedicated Physical Data Channel* (DPDCH). Similar to the downlink, different data rates are realized by using different spreading factors for the DPDCH. Coherent demodulation is used also in the uplink, which requires a channel estimate. Unlike the downlink, where a common pilot signal is used, uplink transmissions originate from different locations. Therefore, a common pilot cannot be used and each user must have a separate pilot signal. This is carried on the *Dedicated Physical Control Channel*, DPCCH. The DPCCH also carries information about the transport format of the data transmitted on the DPDCH. This information is required by the physical layer in the NodeB for proper demodulation.

User-specific scrambling is used in the WCDMA uplink and the channelization codes are only used to separate different physical channels from the same terminal. The same set of channelization codes can therefore be used by multiple UEs. As the transmissions from different terminals are not time synchronized, separation of different terminals by the use of OVSF codes is not possible. Hence, the uplink is said to be non-orthogonal and different users' transmissions will interfere with each other.

The fact that the uplink is non-orthogonal implies that fast closed-loop power control is an essential feature of WCDMA. With fast closed-loop power control, the

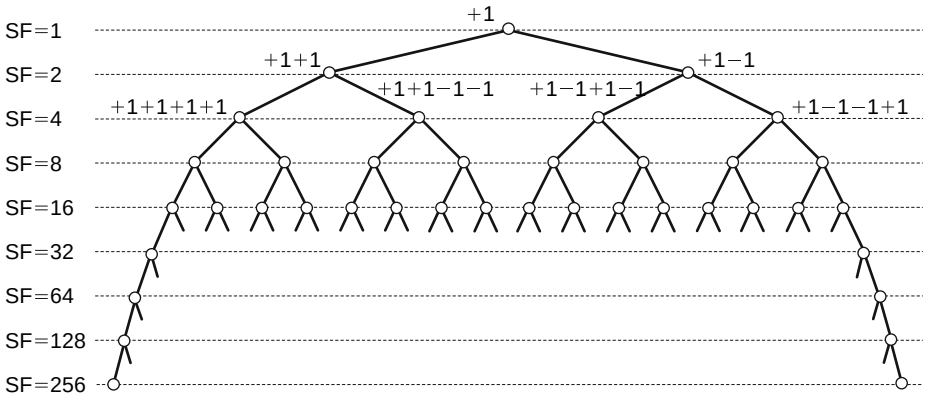


Figure 4.6-5 Channelization codes.

NodeB measures the received signal-to-interference ratio (SIR) on the DPCCH received from each terminal and commands the terminals 1500 times per second to adjust its transmission power accordingly. The target of the power control is to ensure that the received SIR of the DPCCH is at an appropriate level for each user. Note that the required SIR depends on the data rate. This SIR target may very well be different for different users. If the SIR is below the target, that is too low for proper demodulation, the NodeB commands the UE to increase its transmission power. Similarly, if the received SIR is above the target and unnecessarily high, the UE is instructed to lower its transmission power. If power control would not be present, the inter-user interference could cause the transmissions from some users to be non-decodable. This is often referred to as the *near-far problem* – transmissions from users near the base station will be received at a significantly higher power level than transmissions from users far from the base station, making demodulation of far users impossible unless (closed-loop) power control is used.

Fast closed-loop power control is used in the downlink as well, although not primarily motivated by the near-far problem as the downlink is orthogonal. However, it is still useful to combat the fast fading by varying the transmission power; when the channel conditions are favorable, less transmission power is used and vice versa. This typically results in a lower average transmitted power than for the corresponding non-power-controlled case, thereby reducing the average (inter-cell) interference and improving the system capacity.

Soft handover, or *macro diversity* as it is also called, is a key feature of WCDMA. It implies that a terminal is communicating with multiple cells, in the general case with multiple NodeBs, simultaneously, and is mainly used for terminals close to the cell border to improve performance. The set of cells the UE is communicating with is known as the *active set*. The cells that are part of the active set is determined by the RNC based on measurements from the UE.

In the downlink, soft handover implies that data to a UE are transmitted simultaneously from multiple cells. This provides diversity against fast fading - the likelihood that the instantaneous channel conditions from all cells simultaneously are disadvantageous is lower the larger the number of cells in the active set.

Uplink soft handover implies the transmission from the UE are received in multiple cells. The cells can often belong to different NodeBs. Reception of data at multiple locations is fundamentally beneficial as it provides diversity against fast fading. In case the cells receiving the uplink transmission are located in the same NodeB, combining of the received signals occurs in the physical-layer receiver. Typically, a RAKE receiver is used for this purpose, although other receiver structures are also possible. However, if the cells belong to different NodeBs, combining cannot take place in the RAKE receiver. Instead, each NodeB tries to decode the received signal and forwards correctly received data units to the RNC. As long as at least one NodeB received the data correctly, the transmission is successful and the RNC can discard any duplicates in the case multiple NodeBs correctly received the transmission and only forward one copy of each correctly received data unit.[1] Missing data units, in case none of the NodeBs received a correct copy of the data unit, can be detected by the RLC protocol and a retransmission can be requested. Soft handover is one of the main reasons why the RLC is located in the RNC and not in the NodeB.

In addition to reception at multiple cells, uplink soft handover also implies power control from multiple cells. As it is sufficient if the transmission is received in one cell, the UE lowers its transmission power if the power control mechanism in at least one of the cells commands the UE to do so. Only if all the cells request the terminal to increase its transmission power is the power increased. This mechanism, known as or-of-the-downs, ensures the average transmission power to be kept as low as possible. Due to the non-orthogonal property of the uplink, any reduction in average interference directly translates into an increased capacity.

4.6.1.3 Resource handling and packet-data session

With the above discussion in mind, a typical packet-data session in the first release of WCDMA can be briefly outlined. At call setup, when a connection is established between a UE and the radio-access network, the RNC checks the amount of resources the UE needs during the session.

In the downlink, the resources in principle consist of channelization codes and transmission power. Since the RNC is responsible for allocation of resources, it knows the fraction of the channelization code tree not reserved for any user. Measurements in the NodeB provide the RNC with information about the average amount of available transmission power.

In the uplink, thanks to its non-orthogonal nature, there is no channelization code limitation. Instead, the resource consists of the amount of additional interference

[1] Formally, the selection and duplicate removal mechanisms in case of inter-NodeB macro diversity are part of the physical layer. Hence, the physical layer spans both the NodeB and the RNC.

the cell can tolerate. To quantify this, the term *noise rise* or *rise-over-thermal* is often used when discussing uplink operation. Noise rise, defined as $(I_0 + N_0)/N_0$, where I_0 and N_0 are the power spectral densities due to uplink transmissions and background noise, respectively, is a measure of the increase in interference in the cell due to the transmission activity. For example, 0 dB noise rise indicates an unloaded system (no interference from other users) and 3 dB noise rise implies a power spectral density due to uplink transmission equal to the noise spectral density. Although noise rise as such is not of major interest, it has a close relation to coverage and uplink load. A large noise rise would result in loss of coverage of some channels – a terminal may not have sufficient transmission power available to reach the required E_b/N_0 at the base station. Therefore, the radio resource control in the RNC must keep the noise rise within certain limits. The NodeB provides uplink measurements which enables the RNC to estimate the uplink load.

Provided there are sufficient resources in both uplink and downlink, the RNC admits the UE into the cell and configures a dedicated (physical) channel in each direction. Several parameters are configured, one of them being the set of transport formats the MAC layer in the UE and the RNC are allowed to select from for uplink and downlink transmissions, respectively. Hence, the RNC reserves resources corresponding to the highest data rate the UE may transmit with during the call. In the downlink, a fraction of the code tree needs to be reserved corresponding to the smallest spreading factor that may be required during the session. Similarly, in the uplink, the RNC must ensure that the maximum interference in the cell is not exceeded even if the UE transmits at its highest data rate.

During the packet-data call, the data rate of the transmission may vary, depending on the traffic pattern. As a consequence, the transmission power will vary depending on the instantaneous data rate. However, the amount of the code tree allocated for a certain user in the downlink remains constant throughout the packet call (unless the parameters are reconfigured). Furthermore, as downlink transport format selection is located in the RNC, it is not aware of the instantaneous power consumption in the NodeB. The RRC must therefore use a certain margin when admitting users into the system to ensure that sufficient transmission power is available for any transport format the MAC may select. Similarly, in the uplink, the RRC can only observe the average uplink interference and must therefore have sufficient margins to handle the case of all UEs suddenly selecting the transport format corresponding to the highest data rate and therefore the highest transmission power. In other words, the amount of resources reserved for a user is fairly static during the period of the packet call. This is suitable for services with a relatively constant data rate, for example voice services or interactive video transmission, but, significant enhancements are possible for packet-data services with rapidly varying resource requirements.

This page intentionally left blank

Chapter 4.7

4.7

Propagation modelling and channel characterisation

Pierre Degauque, Emmanuel Van Lil, Vittorio Degli-Esposito, Pertti Vainikainen, Filipe Cardosa, Wim A. Th. Kotterman, and Luis M. Correia

4.7.1 Introduction

The development and application of new signal processing algorithms and, more generally, the optimisation of mobile communication systems are strongly related to the statistical properties of the propagation channel. New coding techniques, such as space-time coding algorithms, have been developed as a result of the fast growth in indoor communications coupled with a never-ending need for higher and higher data transmission rates. At the time that the previous COST 273 Action ended, the existing knowledge of channel characteristics was essentially based on data pertaining to path loss, delay spread and coherence bandwidth. However, because this knowledge proved to be incomplete for new systems, intensive research has been conducted concerning the deterministic propagation modelling and the estimation of channel parameters based on actual measurements.

The approach used for electromagnetic propagation modelling depends on the zone under study. For instance, if only the immediate vicinity of the antenna is to be considered, an integral method can be employed, while the coverage area of a transmitting antenna is usually determined via a ray method. Section 4.7.2 describes several recently proposed improvements to various existing techniques: integral methods, differential equation methods, parabolic equations, diffraction theory and Gaussian beams. If, as in the example mentioned above, the ray theory is applied to determine the coverage area of a BS, a large database is needed in order to accurately model the environment's electromagnetic and geometric characteristics. Section 4.7.3 discusses such input database issues, with an emphasis on the complexity that is often cited as the principal obstacle to the widespread adoption of deterministic propagation models. Since decreasing the computation time of ray models is also an important issue, techniques for speeding up ray launching models have also been developed. Simplified or hybrid models, which introduce statistical elements into the deterministic approach, have also been proposed, although they may produce less precise results.

Examples of applications are used throughout the chapter to highlight the advantages and drawbacks of the different methods. Given an assumption of an ideal and complete description of the environment, the hypothesis of smooth surfaces is often implied. However, in many cases, and especially at high frequencies, the roughness or the size of irregularities of a wall surface is on the same order of magnitude as the wavelength, and thus diffuse scattering must be taken into account. This scattering, which plays a fundamental role in time and angle dispersion, is extremely difficult to model. Thus, part of Subsection 4.7.3.5 deals with recent investigations into this topic, in an effort to incorporate the scattering phenomena into the propagation modelling tools.

Modelling as well as experiments providing complementary information about the channel, channel measurement techniques and parameter estimation are detailed in Section 4.7.4. Since the development of MIMO systems implies obtaining complete information about the field characteristics in the vicinity of both the receiving and transmitting antennas, double-directional channel sounding is usually needed. New channel sounder architectures, including antennas, have been proposed to provide precise angular information on the rays' Direction of Departure (DoD) and Direction of Arrival (DoA) and to ensure a better compromise between measurement accuracy, acquisition time and the number of array elements. To increase the accuracy of the channel transfer matrix's spatial and temporal structure, high-resolution

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2006, Luis M. Correia. Published by Elsevier Ltd 2001. All rights reserved

parameter estimation techniques must be used. Improvements in the dedicated algorithms (Multiple Signal Classification (MUSIC), ESPRIT and Space-Alternating Generalised Expectation (SAGE)) are discussed in Section 4.7.4 and examples of applications are presented. Measurements are then processed to extract the channel's statistical properties.

Section 4.7.5 introduces improved empirical path-loss models that are valid for rural and suburban environments as well as office buildings. It is usually assumed that long-term fading is due to a shadowing process and different approaches have been studied in an attempt to improve the well-known model based on randomly distributed Gaussian variables. The rationale behind another well-established assumption – that long- and short-term fading should be modelled separately – is called into question, and a new model based on multiple scattering is described. Lastly, extending the models to the wideband case, which is necessary when the relative delays of arriving waves are large compared to the symbol duration, is also discussed. Clearly, fading statistics are quite important for optimising the communication scheme and for predicting the performance of the link. Still, because numerous advantages would result if the channel fading behaviour could be accurately forecast, the potential of prediction techniques, algorithms and applications has been analysed. Extensive measurement campaigns have been conducted over the last few years in order to obtain statistical information on temporal and angular dispersion, including the characteristics of clusters and of diffusely scattered power, for a variety of propagation scenarios or environments. The main results and conclusions are outlined in Section 4.7.5.

4.7.2 EM theory and diffuse scattering

This section deals with the progress made in the field of the electromagnetic behaviour of communication systems or parts thereof such as antennas, specially taking into account obstacles: terrain (including forests), (complex) buildings and moving objects. Essentially three kinds of method will be discussed:

1. integral methods (moment methods, physical optics and modal expansion);
2. differential equation methods (Finite Difference Time Domain (FDTD) and parabolic equation);
3. ray methods (tracing/launching and Gaussian beams).

4.7.2.1 Integral methods

This subsection deals with the most accurate methods available in electromagnetics to describe waves. Unfortunately, some of those methods require very large amounts of computer resources (both memory and processing time), when applied to scenarios that are large compared to the wavelength, so that they are not yet within the reach of the telecommunication engineer. The work within this COST Action has tried to reduce this gap.

Full moment methods

Of course those methods have been applied to the computation of structures, which are limited in size, such as (elements of smart) antennas [VeLC02b], [TrSi02], [TrSi03] and [NdHH04]. Here the emphasis was placed more on the optimal design of specific antennas, with as the main goal the optimisation of the discrimination of the waves impinging on those antennas by numerical methods described mostly in Section 4.7.4. Paper [VeLC02b] describes in detail the design of 3D array antennas, which allow signal pick-up and discriminate between those signal 'directions' in an unambiguous way, which is not possible with merely planar arrays. This spherical array composed of 12 capacitively compensated monopoles has been designed with a Galerkin piecewise sinusoidal basis function wire-grid program (Numerical Electromagnetics Code (NEC2)) making use of delta-gap excitations. Therefore, the pentagonal parts of the dodecaeder have been approximated by two concentric pentagons and radially extending wires. The emphasis was on the minimisation of the effects of the feedlines on both network parameters and radiation patterns as well as the coupling between the elements. The realised dodecahedral antenna, Figure 4.7-1, has coupling values below −23 dB, while the simulations predict −22 dB. The matching, however, is not perfectly predicted (−18 dB in the 4900–5100 MHz band, while the program predicts a reflection loss of less than −23 dB).

Papers [TrSi02] and [TrSi03] use another commercial Method of Moments (MoM) solver (WIPL) to compute both coupling and impedances of planar circular arrays of monopoles designed around 5 GHz (4.74 GHz), with two, three or four elements. Here the main purpose was to compute the mean capacity of the MIMO channel. Paper [NdHH04] develops an own MoM Galerkin solver and compares the results of a 40 piecewise linear (triangular) basis functions for a half wavelength dipole with

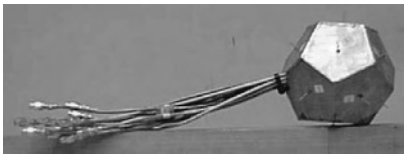


Figure 4.7-1 Practical implementation of the dodecahedral antenna.

the classical sinusoidal distribution. The final goal is the same as in the previous two papers, namely determining the capacity of a (2 × 2) MIMO system.

Moment methods have not only been applied to antennas, which are of limited spatial extent (a few wavelengths in size), but also to the computation of propagation effects, which usually deal with objects that are of significant spatial extent with respect to the wavelength (in fact a few 1000 wavelengths). A lot of work has been devoted in this COST Action to 'correcting' 2D methods to allow computations of 3D structures with nearly the same amount of computational effort as for the 2D case, or even to reduce the effort for the 2D case. The paper dealing with reduction of computational efforts while using full moment methods [KeCu03] is using a Wiener-Hopf technique to find analytical expressions for obtaining the look-up tables for the Combined Field Integral Equation (CFIE). Those are combinations of relatively easy to compute Fresnel integrals. Only for nearly grazing incidence are there some differences between the full integral equation methods and the analytical formulas of the truncated half planes. For a practical case of a German terrain profile computed at 970 MHz, virtually no difference between both methods is seen, even with plate lengths of about 30λ. Another paper [BrCu01] is introducing an integro-differential equation including the gradients of the surface transverse to the direction of propagation. By approximating the phase on the transverse parts of the surface by a second order polynomial, the 2D Magnetic Field Integral Equation (MFIE) for Perfect Electric Conducting (PEC) bodies can be recast into a one-dimensional integral but involving relatively easy to compute Fresnel integrals. This method has been applied to the computation of both ideal theoretical cases like a smooth sloping wedge at 100 MHz and 1 GHz (for both co- and crosspolar fields), as well as to more complex data of real terrains in both Denmark and Ireland. For those cases measured data was available. The conclusions were that the copolar components did agree quite well, but that significant crosspolar components were present, which are justifying accurate modelling of polarisation diversity effects. This is illustrated in Figure 4.7-2.

Physical optics models

Paper [DMSW03] is comparing the standard Kirchhoff formulation with a novel approach called stochastic scattering approach. Here the surface is subdivided in smooth parts with different orientations. Assuming a normally distributed randomness and a Gaussian autocorrelation coefficient, the absolute value of the gradient is Rayleigh distributed with a mean of $(\sqrt{\pi}/L)\sigma_h$, where L is the correlation distance and σ_h is the root mean square of the roughness. While the Kirchhoff formulation leads to scattering patterns that are more dependent for the coherent scattering part from the size of the plane (in this case an $8\times8\lambda^2$ plate is considered), the stochastic scattering approach does show mostly the scattering due to the roughness, which may be much larger than the classical stationary phase assumption in the Kirchhoff case. For this case, with a correlation distance of 4λ and a roughness of 0.35λ, the significant scattering areas are reduced considerably with respect to the Kirchhoff case, which shows a full diffraction pattern. All values are computed at a far-field distance of 2000λ from the plate.

This touches one of the problems of Physical Optics (PO) for large objects or close transmitters or receivers: the plane-wave approximation is not valid any more, leading to fields that continue to grow if the size of the scatterer is increased. Therefore, other solutions have to be found. Paper [LCDC02] is using a parabolic phase

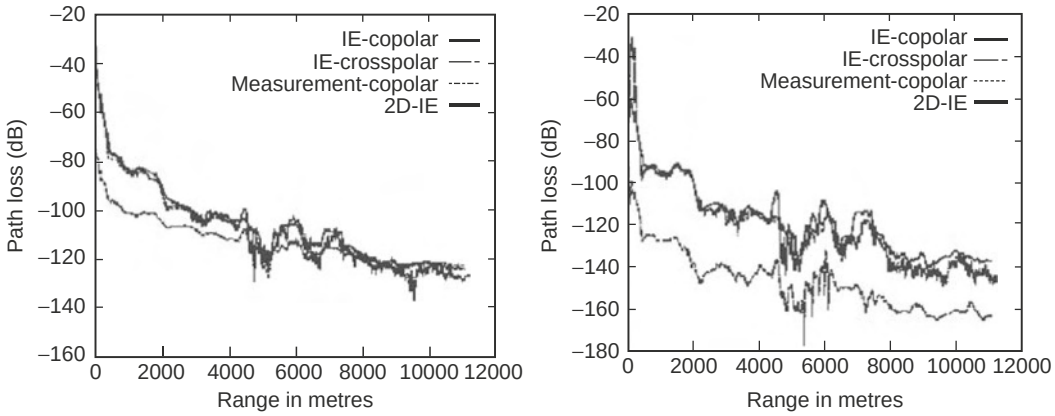


Figure 4.7-2 Co- and crosspolar fields for Danish terrain.

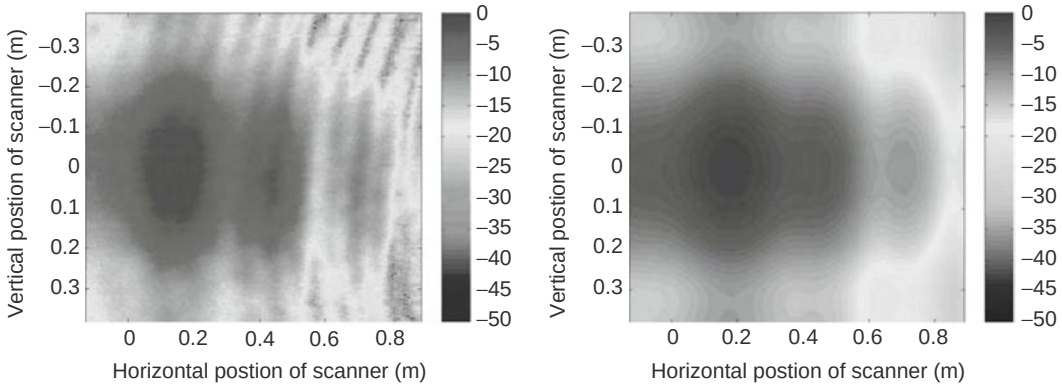


Figure 4.7-3 Amplitude of the reflected field of a 0.35 × 0.5 × 0.02 m Plexiglass plate at 5.8 GHz.

approximation for this purpose, which has been developed in a previous COST Action, namely COST Action 259 –Wireless Flexible Personal Communications (COST 259). This paper investigates many further refinements of this Fresnel method such as

1. the use of near-field formulas;
2. using internal reflections in a reflecting plate of finite thickness and material properties (permittivity and permeability);
3. optimal choice of the integration centre, which is critical for the efficient application of this method.

For this last case, seven different methods (starting with the trivial solution) have been investigated. It was shown that the optimally fitted method (optimally means that the parabolic phase behaviour is fitted to osculate the real spherical one as efficiently as possible, i.e. up to and including second degree terms), gave excellent results with only four times the number of FLOating Point operations per Second (FLOPS) of the trivial solution. Both simulations and measurements in an anechoic room on a flat plate also agree very well, both in amplitude, Figure 4.7-3, and in phase, Figure 4.7-4.

Modal expansion models

Since this method is specially suited for closed areas, one contribution [Kyri03] has applied this method for the evaluation of the K-factor of a hallway. Indeed, the Ray-Tracing (RT) methods are less appropriate to this kind of problem, since the number of reflections on the boundaries of the structure may become very high. The contribution of many different propagating modes is evaluated, taking into account the losses (both dielectric and conduction) with an equivalent reflection coefficient. The constant component estimation proved the most stable, better than the statistical method or the ESPRIT algorithm. The intuitively expected behaviour is confirmed, namely that the K-factor increases with increasing losses, and with decreasing size.

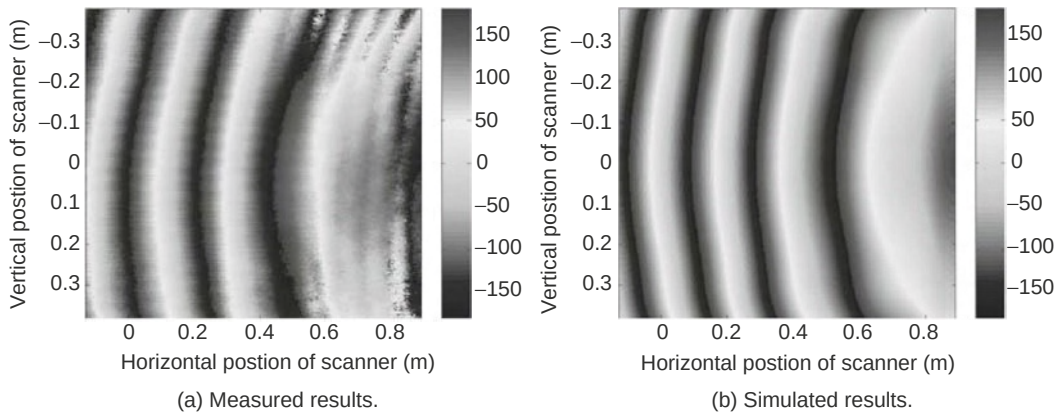


Figure 4.7-4 Phase of the reflected field of a 0.35 × 0.5 × 0.02 m Plexiglass plate at 5.8 GHz.

4.7.2.2 Differential equation methods

FDTD models

Method to terminate the calculation space in FDTD in a computationally efficient manner

The ideal truncation method has a low data storage requirement, an easy construction procedure, high flexibility, high stability and high accuracy. No current truncation method satisfies all of these points. A method which shows better accuracy may experience lower stability. Some methods such as the perfectly matched layer are especially good at gaining accuracy and some methods such as Higdon, Mur 1st and 2nd, and Liao's Absorbing Boundary Condition (AbBC) are good at gaining flexibility and have simple construction procedures with relatively small data storage requirements. Liao's AbBC is interesting in that it does not have any media parameter such as relative permittivity. This means Liao's AbBC can be easily constructed in the inhomogeneous environment, while other boundary conditions based on one-way wave equations such as Mur 1st and Mur 2nd AbBCs have media parameters. The nature of Liao's AbBC has been previously investigated [RaAr95], [ChWa92], [PrSh97], and [Rama99] concludes that Liao's higher order AbBC is unstable and even a 2nd order Liao's AbBC exhibits some extent of instability. The higher order Liao's AbBC also requires more memory to store the data for the update of the electromagnetic field values at the boundary. [Cost03] therefore has studied Liao's second order boundary condition with an emphasis on obtaining a reduced memory requirement and improved stability. The results of this work can, however, be applied to higher order Liao's AbBCs. The detail of the numerical formulation is given in [Cost03]. This novel AbBC reduces the data storage requirement for the boundary calculation of the original AbBC by 33%. [WaCh95] pointed out that the high stability is realised when the poles of the artificial reflection coefficient are within the unit circle of the complex plane. Based on this, the possible instability caused by the proposed boundary condition is investigated, compared with the original 2nd order Liao AbBC. The study showed another merit of the proposed AbBC, that is, the proposed AbBC gains its stability with $\alpha \leq 0.5$ while the original AbBC has instability with any α. The reflection coefficient from the proposed AbBC in the time domain was calculated with the marginally stable damping parameter $\alpha = 0.5$, compared with Mur 1st and 2nd AbBCs which are originated from one-way wave equation like Liao's AbBC. The artificial reflection coefficients from the proposed AbBC and Mur 1st AbBC were in a comparable order, while the Mur 2nd AbBC showed the best absorption among these three boundary conditions. On the other hand, the analysis of the frequency reflection coefficient proved that the proposed method was a better uniform absorber than Mur 2nd AbBC, which is another merit of the proposed AbBC.

Method to speed up the Frequency Dependent (FD) FDTD simulation which accommodates the frequency dependent media for UWB (Ultra Wide Band) systems by removing the Courant–Friedrichs–Lewy (CFL) stability condition

In recent years, the Alternating Direction Implicit (ADI) method [Smit65] has been introduced to the FDTD [Nami99], [ZhCZ00]. The standard explicit FDTD [Yee66] has to satisfy the CFL bound [Tafl95] to be stable. Cases requiring a small spatial discretisation lead to excessive time sampling and increase the total Central Processing Unit (CPU) time required for the FDTD. The maximum Δt possible in FDTD under the CFL condition is defined as Δt_{CFL}. On the other hand, ADI-FDTD is unconditionally stable when taking sampling times Δt greater than the limit, keeping the total elapsed time within a practical range. $\Delta t/\Delta t_{CFL}$ is called the Courant-Friedrichs-Lewy Number (CFLN) N_{CFL}. Thus ADI-FDTD is attractive when, for example, studying the influence of skin effects in conductive materials or signal reflections in lossy media for microwave imaging with UWB systems. These UWB systems typically have frequency dependent lossy media parameters but ADI-FDTD and FDTD have media parameters which are constant across frequencies. The adaptation of ADI-FDTD to such dispersive environments is required to accommodate UWB systems. Holloway *et al.* [SHBPM03], [HMDA02] accounted for material dispersion in the magnetic field and mentioned the possibility of dealing with material dispersion in the electric field. However, [SHBPM03] and [HMDA02] do not give an implementation of ADI-FDTD which deals with material with FD permittivity ε and ohmic losses. Yuan and Chen's [YuCh03] concern focuses on conductive materials. Garcia *et al.* [GRBM03] deal with the Debye model utilising the electrical polarisation vector field. The COST 273 project [Cost04], [ThCo04], [CoTh04a], proposes an alternative technique using the electric flux density $\boldsymbol{D}$, similar to the work by Schmidt and Lazzi [ScLa03], but with FD materials for UWB systems. This runs parallel to [GRBM03] and provides convergent evidence for the general approach to FD materials in ADI-FDTD class problems. The novelty of this work is in proposing an FD-ADI-FDTD which can deal with Debye dispersive materials with ohmic losses, along with the solutions to the problems on the $\boldsymbol{D}$ source excitation. [CoTh04a] detailed a method of dealing with the CFL stability condition in a lossy medium for UWB systems, utilising the frequency spectrum of the source excitation. The error in the lossy case was proved to equal the lossless case with an appropriate N_{CFL} adjustment. It is

shown that FD-ADI-FDTD is applicable for problems which can accommodate error higher than 0.1 with $N_{CFL} \geq 6$, gaining a simulation time benefit over FD-FDTD. Although the merit of FD-ADI-FDTD is the absence of a limitation on the value of Δt for stability, which allows the reduction of elapsed time of FD-ADI-FDTD, compared with FD-FDTD, the accuracy reduces this merit by a certain amount because accuracy largely depends on the temporal discretisation and spatial discretisation. Therefore, a method to set an appropriate Δt in UWB systems is also inductively proposed [CoTh04b].

Parabolic equation methods

The Parabolic Equation (PE) method is based on a paraxial approximation to the wave equation. Assuming that the fields propagate in a preferred propagation direction, this is a full-wave method accounting simultaneously and accurately for all wave phenomena like diffraction, refraction and scattering. Provided the field is known on an initial plane and adequate boundary conditions on the scattering objects and at the outer boundaries of the integration domain are given, it is easy to proceed (march) in range in the propagation direction. Here, no heavy integrals and matrices are involved. This explains the success of this method for propagation purposes, specially if the equation is reduced to its 2D version. It is very accurate at angles within 15° of this propagation direction. Several contributions apply the 2D version of this method to ducting phenomena on microwave propagation. Indeed, ducting is present in 15% of the time all over the world, specially over humid environments like coastal zones, and hence being able to compute the propagation in those circumstances is important. In [SiMi03] it is shown that surface-based ducts affect the propagation in short distances (up to 3 km) by provoking the shift of the location of the interference

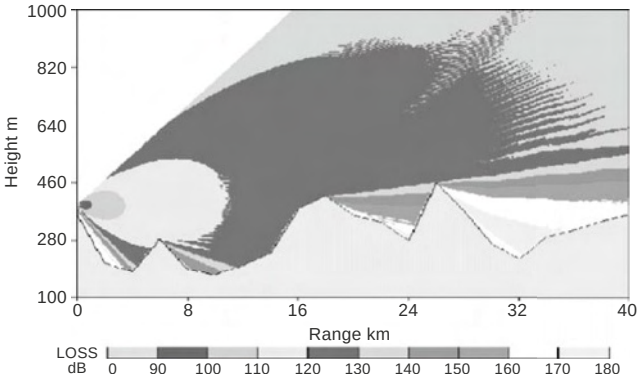


Figure 4.7-5 Path loss for standard troposphere conditions.

minima. At those minima, the differences between the case with the standard troposphere and the ducting case may exceed 20 dB for a ducting layer of 50 m thickness. Paper [SiMi03] discusses the influence of the changes in the evaporation duct thickness on two links at 2.5 and 5.8 GHz. Examples of range dependent ducting (corresponding to refractivity profiles along a mixed land-sea path) are also reported. For this last case between 5 and 10 dB variation in the losses can be accounted for. In both above-mentioned papers a smooth perfectly conducting underlying surface is assumed. The influence of the terrain is introduced in [SiMi04] and [SiMi05]. Assessment of the combined effect of terrain and ducting on UMTS path losses at UMTS frequencies (2.1674 GHz) is reported in [SiMi04]. Ducting can help to increase signal in valleys. This is seen when comparing the standard case, Figure 4.7-5, with the ducting case, Figure 4.7-6. Relying on the precision of the PE and its ability to provide quantitative assessment of the losses, in [SiMi05] the importance of possible inaccuracies in the terrain profile data for the Sofia region is discussed. The influence of the

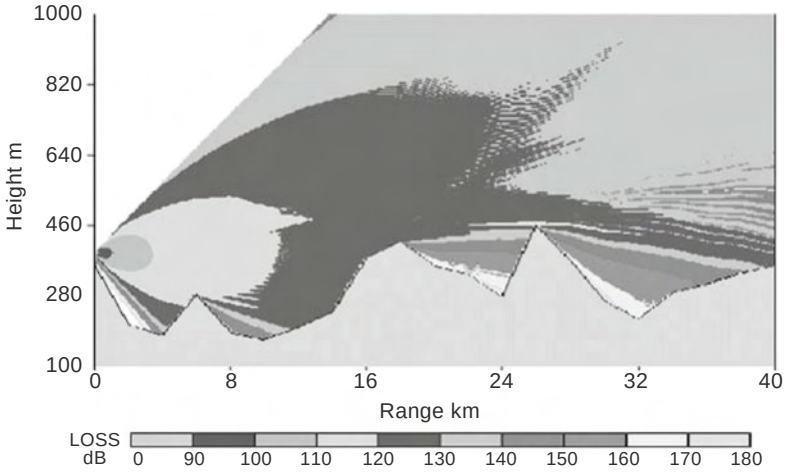


Figure 4.7-6 Path loss under a 100 m ducting layer between 450 and 550 m.

ground type (for vertical polarisation) is also discussed. Another kind of problem that is very well suited for solutions based on PE methods is the case of wave propagation over a forest edge. The results for horizontally polarised links at 1.3555 and 1.5995 GHz are compared in [WaHo03] with Geometric Theory of Diffraction (GTD) approximations, which are in general up to 20 dB worse that the measured cases for low receiver heights. The 2D PE solution with a Leontovitch boundary condition for the ground and an absorbing layer above the region of interest agrees in general within ±5 dB with the measured values.

4.7.2.3 Ray methods

Here two fundamentally different methods can be distinguished. One is using beams that are not localised in space, but that occupy the whole extent of space (and hence satisfy Maxwell's postulates). The Gaussian beams are one possible implementation. This will be discussed in the first subsection. The second uses the concept of rays, which are cuts of phenomena that do satisfy as a whole Maxwell's equations (plane, cylindrical or spherical waves). Some methods do trace the rays completely and compute the different reflection and diffraction points, which is a cumbersome geometric procedure. Some others do launch rays, and use a receiving sphere to determine the ones relevant to the actual receiver position.

Gaussian beams

Paper [TaLH02] uses a Beam Launching (BL) method to compute the fields in indoor environments. One first decomposes the source fields into a sum of beam field by the Gabor frame decompositions. This procedure is fully equivalent to the decomposition of antenna fields into plane or cylindrical waves. The expressions for reflection of a Gaussian beam on a plane boundary is then derived, and applied to computations of indoor channels at mm wave frequencies (57–59 GHz). Indeed at those frequencies, the diffractions are negligible and hence no diffraction is taken into account. Measurements in a small laboratory confirm the validity of the procedure, even if only the dominant path is predicted accurately.

Geometrical optics (GO) models

Many papers are treating this subject and provide applications in the world of (mobile) communications. Some just use Uniform Theory of Diffraction (UTD)/PO for further investigations, which will be described in more detail in Section 4.7.3, but some do provide treatments of specific wave phenomena like diffraction and will be discussed here. Paper [GLBM02] uses GO to compute the effects of wind turbines on communication and safety systems (such as radars). Different effects, depending on the distance from the turbines to the system under investigation, are noticed:

1. a deep shadowing, but of limited geographic extent, turning into a smooth shadowing that reduces the range of the system. This effect decreases considerably with decreasing frequency and is hence only important for radar systems;
2. zones of occurrence of false echoes (ghost images) for radars.

In the same paper, the delay spreads for a communication system as well as Resource Description Framework (RDF) accuracy are also discussed. Both turned out to be negligible for the turbine parks investigated and for the system considered, but might become important if the turbines are placed too closely to the systems. Finally, the effect of a turbine park on a Differential Global Positioning System (DGPS) system (working at 312 kHz) is investigated with a moment method. It turned out that the transmitted wave towards the sea was even improved for this particular park, which was acting as parasitic antenna at that low frequency. The three-dimensional extent of the areas, where the four different kinds of false echo can occur, were further quantified with spherical-wave approximations in [LiTC05]. Paper [HHBW01] uses a simple ray-optical prediction tool to evaluate the path loss in urban areas. This information is used for UMTS load planning. One of the few papers that deal with extensions of the classical UTD theory is paper [ElVa03]. Here, the theory for a lossy dielectric wedge is derived, by taking into account both the angle of incidence and diffraction into the multiplication factors. This improves significantly the diffracted fields for lossy dielectric wedges over just using the plane wave Fresnel reflection coefficients.

Another [TrLC04] deals with the implementation of curved surfaces in classical 3D RT programs. The most difficult part is not so much the electromagnetic treatment of the diffraction, but to find a fast way to determine the reflection point of a wave on this curved surface, in this case a cylinder. An iteration process converging in all special cases within six steps within an accuracy of 0.1% of the transmitter–receiver distance. Paper [RoMGPJL04] deals with a solution for multiple edge diffraction with a spherical incident wave for antenna location above, level and below the knife edge plane. The results are in good agreement with the Xia-Bertoni results, while requiring much less computer time, and are accurate within at most 2.27 dB. A general program using a full 3D GTD theory, but tuned to MIMO analysis, is described in [GaHS03]. Here, the efficiency is increased by simplifying the geometry to a discrete number of single-bounce reflectors, called scatterers.

Comparison with real measurements at 1.95 GHz and a bandwidth of 120 MHz yield good accuracy for the single-bounce echoes in a half-open environment, just by tuning the size of the scattering clusters, the number of scattering coefficients and the Rician K-factor.

RT methods can also be used to determine the permittivity of building materials. Indeed, [JKVW05] uses simple amplitude-only reflection measurements at different incidence angles to fit the permittivity of flat walls. It is even possible to introduce diffuse radiation into the reflection model, giving a more realistic behaviour of the scattered field, by introducing a scattering coefficient, which increases as expected from a small value (0.05) for a nearly flat hangar wall, up to 0.4 for more irregular rural building walls [EFVG05].

4.7.3 Deterministic propagation modelling

4.7.3.1 Introduction

Deterministic propagation modelling is aimed at reproducing or studying the actual physical radio propagation process for a given environment. Such approach is particularly suitable for manmade environments, where radio waves interact with relatively simple, geometric obstacles such as buildings and streets. In this case, the geometric and electromagnetic characteristics of the environment and of the radio link can be described and stored in files (databases) and the corresponding propagating field can be computed through analytical formulas and/or computer programs. The degree of accuracy and of 'determinism' varies depending on the chosen approach. While some approaches can be properly considered 'deterministic' others may be defined 'hybrid' or 'semi-statistical', as long as they leave off some deterministic aspects and/or describe them through average, statistical parameters, thus decreasing complexity and computation time. The degree of determinism is also related to the accuracy of environment representation. For example, the adoption of a very accurate, full wave FDTD method in an environment described in terms of soil usage and building density would be a nonsense. Commercially available urban databases are not very accurate and detailed. Therefore, some propagation effects which are critically sensitive to database accuracy (e.g. fast fading, diffuse scattering, phase, polarisation, etc.) must be necessarily neglected or treated in a statistical way. Deterministic propagation models have advantages and drawbacks with respect to statistical models such as Hata-like models or statistical channel models. Deterministic models are physically meaningful, accurate and flexible. Statistical models on the other hand are simpler, faster and more synthetic because their output can describe an entire class of problems instead of a single one.

Generally, statistical channel models are used in the link level design phase while deterministic models are used in the planning phase of mobile radio systems. However, despite the fact that the use of deterministic field prediction may in principle lead to great advantages in terms of deployment cost reduction and service quality increase, the widespread adoption of deterministic models is still very limited due to the high cost and the low reliability of input databases and to the high complexity and computation time of the corresponding computer programs. For these reasons, most research within COST 273 addressed topics such as database handling [ScWi03], [CoLA03], [DeFA04], sensitivity to database accuracy [NaCB04] and CPU time reduction [HoWW03], [CoLA03], [DeFA04], [WWWW04], [BCFF02]. In principle, deterministic models such as RT allow a multidimensional characterisation of propagation, i.e. may give information on the distribution in time and space (Angle of Departure (AOD)/Arrival(AOA)) of the multipath field. This may be very valuable for the design and planning of present and future systems using array antennas and/or space-time coding techniques such as MIMO systems. However, due to input database inaccuracies and limitations in the modelling of diffuse scattering and of the multipath propagation process in general, multidimensional performances of today's models are still quite poor. Some authors have therefore worked on the improvement of deterministic modelling capabilities [DDGW03], [DEGDA04], [DEESGK02a], [FuMa03], [BHHT04a], [JePK05], [GhTI04], [Bert02]. Finally, deterministic models and their realistic propagation simulation capabilities have been applied to system assessments such as network planning issues, system performance assessments, positioning techniques etc. [CoWS04], [WHZL02], [ZBLL04], [Fusc04]. Deterministic and semi-deterministic field prediction models and techniques, their performance, their pros and cons are described in this section. Applicative aspects and comparative overview of these models include such issues as database accuracy sensitivity, computation time, diffuse scattering modelling, and performance metrics. Since reliability and cost of input databases represent major concerns, Subsection 4.7.3.2 is fully devoted to this topic. The various models are then treated starting from the more rigourous, complex (4.7.3.3) to the more simplified ones (4.7.3.7).

4.7.3.2 Input database issues

Deterministic propagation models require as input a proper description of the environment. In the case of urban propagation modelling, which is of interest here, both a geometrical and an electromagnetic description

of each object (building, wall, terrain) must be provided and stored in the database. Outdoor databases can be extracted from aerial views using stereo-photogrammetry, from digital cadastre data or by digitising analogue maps. If the space domain is discretised into pixels, and for each pixel a geometrical or electromagnetic parameter is stored, then a *raster* database is obtained. Otherwise, if the environment is decomposed into basic geometrical elements (triangles, polygons, segments, etc.) then a *vectorial* database is obtained. Vectorial databases are usually more compact and accurate than raster databases and are therefore adopted by default in the present section, except where otherwise stated. In vectorial databases, buildings are usually represented as polygonal prisms with flat tops. In all cases, the intrinsic cost of urban databases is high, ranging from some hundred to some thousand euros per square kilometre. Moreover, the attainable accuracy is necessarily limited. The standard deviation of the error in urban databases is of the order of 0.5 m for horizontal coordinates and much higher for vertical coordinates. Vertical dimensions are often unavailable or particularly subject to errors, due to deviations in rooftops from flat polygons, with an evident impact on the reliability of *over-rooftop* propagation models. Moreover, environment cluttering data (vegetation, vehicles, street signs, etc.) are often unavailable or impractical to model. While reliability must be considered a very advisable quality, the 'level of detail' is not necessarily so. A great level of detail (sometimes confused with 'accuracy') means the presence of a great deal of building elements (internal yards, indentations, edges, roof structures, etc.) which, although relatively irrelevant, can still greatly increase the size of the database, with a more than proportional increase in the computation time of the prediction. Therefore, geometric databases must often be properly 'purged' or simplified before they can be input to deterministic propagation models.

In the indoor case, the description of floor and room partitions of the buildings in AutoCAD's native file format for CAD models (DWG) or Data Exchange Format (DXF) can often be extracted from the architectural design files of the building. In this case the accuracy of the database can be good. However, furniture is of course not included: the reliability of the field prediction with respect to measurements on a real, furnished building is therefore questionable no matter the accuracy of the adopted model. The electromagnetic description of both outdoor and indoor environments is also inaccurate because of the lack of related information. Compound materials are often treated adopting *effective* electromagnetic parameters, i.e. parameters representative of the overall behaviour of the compound in terms of reflection, transmission and diffraction of an incoming wave. A usual choice is to use literature or empirical values for the permittivity (ε_r) and the conductivity (σ) of each class of materials. For example, $\varepsilon_r = 5$ and $\sigma = 10^{-2}$ are usually adopted for typical building walls in European cities.

Each environment type (indoor, dense urban, suburban, etc.) has appropriate database formats. Each database format is associated to appropriate propagation models. Therefore mixed environments pose a number of problems to both database handling and field prediction modelling. The different types of data must be harmonised by means of proper conversion, interpolation, over- or undersampling and/or the different propagation models must be somehow combined. Although few authors within COST 273 worked on database issues, it is evident that problems such as high cost, low reliability and high complexity of input databases (and the corresponding high handling and prediction time) probably represent the main limits to the widespread adoption of deterministic propagation models. Studies related to database extraction, handling and simplification carried out within COST 273 are briefly reported in the remaining part of this subsection.

Electromagnetic parameters determination

In [ScWi03], RT simulation of wave propagation in hospitals in the frequency range from 600 MHz to 5.2 GHz is considered. What makes wave propagation in hospitals different from that in other buildings is the special construction of their walls. Some of these walls contain metallic layers, e.g. operating rooms with CrNi-steel faced walls, X-ray rooms with lead shielded walls or Magnetic Resonance Tomography (MRTo) rooms with Electro-Magnetic Compatibility (EMC) shielded walls made of a copper foil. If the real conductivity of those metals (copper, lead, CrNi-steel) were used in the simulations, this would lead to an infinite shielding of the walls, because slots and openings in the metallic layers cannot be taken into account. For this reason, the *effective* constitutive parameters of each kind of wall were determined through multiple measurements of single wall attenuation in different hospitals and double checked by comparison between measurement and simulation of attenuation between two adjacent rooms. The attenuation of single walls was measured by positioning horn antennas on both sides of the wall and measuring the attenuation of the transmitted power directly through the wall. For most of the walls it was possible to obtain constitutive parameters that are valid over the whole frequency range from 600 MHz to 5.2 GHz (see Table 4.7-1). However, this was not possible for the walls of the MRTo room, because the copper EMC shielding of this room shows a sensible frequency dependent behaviour. Therefore, its effective conductivity was determined separately for all five frequencies, Table 4.7-2.

Similar measurements on electromagnetic parameters were carried out by [JePK05] and [JKVW05] by

Table 4.7-1 Composition of walls in hospitals and constitutive parameters.

Type of room	Material	Thickness	ε'	ε''	σ(*S/m*)
Sick room	Plasterboard	2.5 cm	2.4	0.01	0.0
	Rock wool	7.5 cm	1.2	0.02	0.03
	Plasterboard	2.5 cm	2.4	0.01	0.0
X-ray room	Plasterboard	2.5 cm	2.4	0.01	0.0
	Lead	1.0 mm	1.0	0.0	70
	Rock wool	7.5 cm	1.2	0.02	0.03
	Plasterboard	2.5 cm	2.4	0.01	0.0
Operating room I	CrNi-steel	0.8 mm	1.0	0.0	1.0
	Plasterboard	1.8 cm	2.4	0.01	0.0
	Rock wool	7.5 cm	1.2	0.02	0.03
	Plasterboard	1.8 cm	2.4	0.01	0.0
	CrNi-steel	0.8 mm	1.0	0.0	1.0
Operating room II	*Trespa*™	0.5 cm	3.0	0.1	0.0
	Plasterboard	2.0 cm	2.4	0.01	0.0
	Lead	0.5 mm	1.0	0.0	20
	Rock wool	7.5 cm	1.2	0.02	0.03
	Plasterboard	2.0 cm	2.4	0.01	0.0
	Trespa™	0.5 cm	3.0	0.1	0.0
MRTo room	Chipboard	1.3 cm	2.5	0.1	0.0
	Styrofoam	2.0 cm	1.1	0.0	0.0
	Copper shield	0.16 cm	1.0	0.0	σ *(f)* (see Table 4.7-2)
	Rock wool	7.0 cm	1.2	0.02	0.03
	Plasterboard	2.5 cm	2.4	0.01	0.0
	Rock wool	7.5 cm	1.2	0.02	0.03
	Plasterboard	2.5 cm	2.4	0.01	0.0

performing both reflected and transmitted power measurements on different kinds of walls with directive antennas. An IEEE 802.11b transmission system between two Power Controls (PCs) was used. In order to separate the desired reflected or transmitted signal from spurious signals coming from surrounding objects, a time-domain resolution method using the Pseudo-Noise (PN) sequences of the IEEE 802.11b was adopted. The electromagnetic characteristics were then determined by a 'feedback' iterative technique where the Electro-Magnetic (EM) parameters in the input to a 3D RT (see Subsection 4.7.3.4) were varied until the measured value was reproduced with sufficient accuracy (see Figure 4.7-7).

Mixed database handling

A variety of radio communication systems such as DAB and DVB, 3G public networks and Wireless Local Area Networks (WLANs) must coexist in urban and suburban environments. To assure high-quality in-building coverage for the different systems in different environments (indoor, microcellular, macrocellular, etc.), or to control the limits and interferences between the various cell layers in a complex network, deterministic simulation tools are required in conjunction with appropriate multi-environment database handling techniques. In [LoCo02a], [CoLA03] and [LoCo05] a multi-environment prediction tool based on a 2.5 ray-launching technique for outdoor and for outdoor–indoor propagation and a 3D–RT for indoor propagation is presented (see also Subsection 4.7.3.4). An optimised indoor database format was adopted to speed up indoor prediction without significant loss in accuracy. Only the main structures that significantly infer on the propagation mechanisms (typically the floor surfaces, the facades and the concrete partitions) are used in the computation of the multipath ray trajectories, the light partitions are only taken into

Table 4.7-2 Frequency dependent conductivity σ *(f)* of EMC shielded MRTo-rooms for RT simulations.

Frequency (MHz)	σ (*S/m*)
600	750×10^2
900	400×10^2
1800	230×10^2
2450	120×10^2
5200	70×10^2

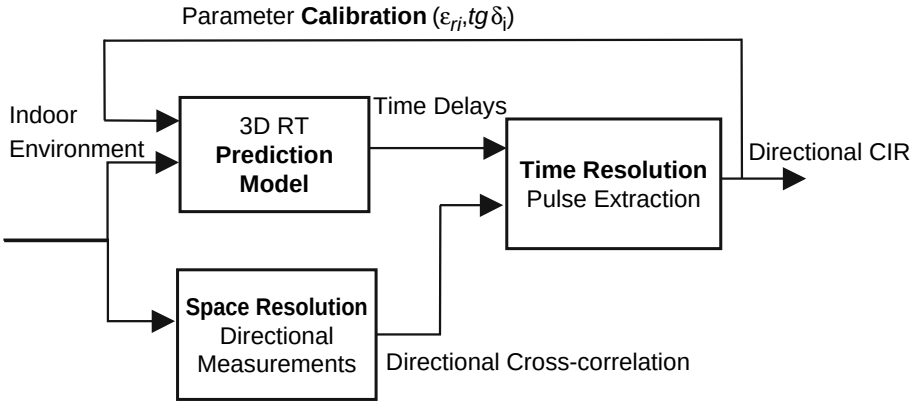


Figure 4.7-7 Iterative technique for EM parameter determination.

account to determine path attenuation. The distinction between the main structures and the light partitions is incorporated in the building database. The analysis of the hybrid radio networks (coverage and QoS) requires the outdoor and indoor predictions to be done in a unique geographical system. Hence both kinds of terrain modelling, the outdoor map data and the 3D building representation have to be perfectly adjusted. Land usage data is also considered where available. The indoor representation that is derived from architect plans is generally available in a local coordinate system. Successive isometric transformations must be used to adjust the local representation to the building contours in the outdoor map data. Mixed database handling issues are also addressed in [KüMe02] and [KEGJ03]. In the former paper, empirical models for outdoor to indoor penetration are presented and validated with measurements. Of course, since the two different environments must be handled in a uniform way, proper conversion procedures must be adopted. In the latter, an automatic propagation model for the automatic planning of UMTS networks is presented which is one of the results of the European Project Information Society Technologies (IST) Models and Simulations for Network Planning and Control – IST Project (MOMENTUM). Since the model must work in different deployment scenarios (macro-, microcellular and indoor), algorithms to handle transition between areas with digital terrain and urban databases of different resolution have been developed.

Geometric database simplification

Urban field prediction tools are always run over a discrete set of test spots, a route or a limited area such as a cell or a portion of a city. Since computation time strongly depends on the size of the input database, it is necessary to minimise its size by selecting only the *active* area of the topology, i.e. only the buildings or obstacles that participate in the propagation process. As far as field strength prediction is concerned, the active area is limited to the obstacles *around* and *between* the radio terminals, therefore the active area can easily be 'guessed' using rule-of-thumb considerations. On the contrary, when the time and angle distribution of the multipath is to be determined, even far, but prominent, obstacles can have a significant impact on prediction results, and therefore a scientific criterion is needed. A method to attain this aim is presented in [DeFA04]. The method is based on the idea of running a rough, heuristic field prediction model aimed at identifying the 'active' building set. The basic assumption is that paths experiencing a single (or two) interaction(s) with building walls (often referred to as 'first (second) order rays') are likely to be more relevant than rays experiencing multiple interactions, since more interactions correspond to both a higher number of concentrated losses and a higher distributed path loss. Thus, the simplification method is based on the selection of the buildings belonging to three categories (A, B, C) as follows.

(A) buildings located around and between the radio terminals, buildings located within an ellipse of focuses TX and RX, and of given eccentricity are selected. Usually the eccentricity value is chosen so that the terminals are completely surrounded by buildings.

(B) buildings which can be directly 'seen' from either the TX, the RX or both terminals, therefore generating first- or second-order rays. This step is performed by running a simple geometric visibility tool de-embedded from a RT tool.

(C) buildings not belonging to category A and B , but which are likely to have an impact on propagation because of their height (greater than average) or orientation, i.e. generating low-order rays.

Referring to the ellipse mentioned in (A), its centre is chosen as the origin O of a cylindrical reference system and the map is divided into angular sectors of given amplitude

$\Delta\Psi$. For each sector the average building height H_a is computed and a height tolerance Δh is set. Then, all buildings well above average height (with $H > H_a + \Delta h$) are selected. Buildings of average height (i.e. with $H_a - \Delta h < H < H_a + \Delta h$) are selected only if they have at least one wall almost perpendicular (within a given angular tolerance to the radial coordinate). This means that they are selected if they ideally could back-scatter energy toward the ellipse (hence from the TX to the RX). Once classes A to C are identified, the corresponding buildings are collected into the simplified map and duplicates discarded. If more than one RX (TX) is present (ex. receiver path) the whole algorithm is performed once for every RX (TX) and each time the new buildings are added into the simplified map.

In Figure 4.7-8 a simple case in the city of Helsinki is shown. It is shown in [DeFA04] that in this case the CPU time of an RT prediction can be reduced from 2131 minutes (complete map) to 300 minutes (simplified map) virtually without degradation in DS and AS prediction. On the contrary, wideband results obtained with the guessed map are unacceptable.

4.7.3.3 Electromagnetic models

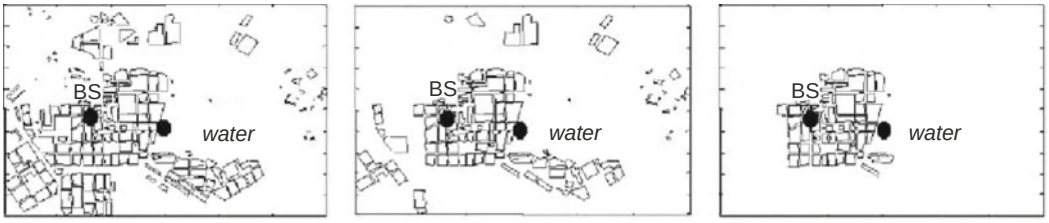
Electromagnetic methods are aimed at directly solving Maxwell's equations in some sort of discretised way (see Section 4.7.2). Depending on the adopted form of Maxwell's equations and on the kind of discretisation of the domain, various methods can be identified such as Finite Element Method (FEM), FDTD, MoM and others. Electromagnetic methods are not commonly used for field prediction at radio frequencies, the main reason being the small wavelength with respect to the dimensional scale of the environment in this kind of problem. In fact, since the space discretisation step must be a fraction of the wavelength, the overall grid dimension would be huge, and consequently memory occupation and computation may become unbearable. Moreover the overall level of accuracy both available in input databases and required in output prediction is usually not so high as to justify the adoption of sophisticated numerical methods. Electromagnetic methods require a detailed description of both the geometry and the electromagnetic parameters of the environment. Real (as opposed to effective) electromagnetic parameters must be used for each single structure composing the environment.

Only FDTD has been used for deterministic field prediction within COST 273. In [Schä03], an FDTD method is used to investigate wave propagation between adjacent rooms in Hospitals in the frequency range from 42.6 MHz to 300 MHz. For an accurate simulation of propagation in such complex environments as hospital rooms, with metallic layers in walls and with structures that are small compared to the wavelength such as slots in metal layers, numerical tools based on the solution of Maxwell's equations and a detailed description of each layer composing each room wall are necessary. Wave propagation between two adjacent rooms is investigated for different frequencies and polarisations and results are compared with measurements. The upper simulation frequency is limited due to the limited Random-Access Memory (RAM) resources of the computer, because the calculating area has to be divided in cuboids that are much smaller than the wavelength. With about 4 giga byte (GB) of RAM and a calculation area of about 10 m × 5 m × 3 m, it is possible to simulate frequencies up to about 500 MHz. In order to determine the so-called 'room attenuation', a transmitter was located inside a room and the pathloss to the adjacent rooms was simulated or measured.

In Figure 4.7-9 room attenuation with vertical polarisation for different kinds of rooms is reported. It is evident that the simulations are in good agreement with the measurements. Moreover, room attenuations for different polarisation are obtained: results show that, due to the orientation of slots inside walls, room attenuation and the field pattern strongly depend on polarisation.

4.7.3.4 Ray models

Ray models are based on GO theory and its extensions to treat reflection and transmission on plane surfaces and diffraction on rectilinear edges (see Section 4.7.2). GO theory is based on the so-called ray approximation, which is valid when the wavelength is sufficiently small compared to the dimensions of the environment obstacles.



(a) Complete map (183 buildings). (b) Simplified map (118 buildings). (c) Guessed map (69 buildings).

Figure 4.7-8 Central Helsinki maps; black dots = radio terminals.

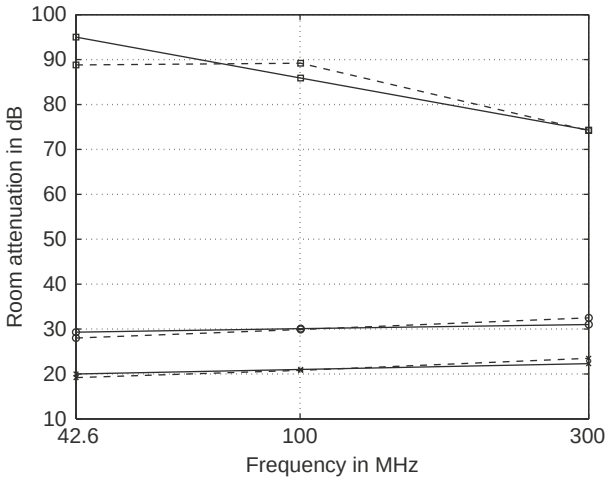


Figure 4.7-9 Comparison of the measured and simulated room attenuations. Drawn through: measurement, dashed: simulation. Squares: MRTo room, circles: operating room, x-marks: X-ray room.

This is usually the case in urban radio propagation problems. Under this assumption, the electromagnetic field and its multipath propagation can be expressed in terms of a set of *beams* or a set of *rays*, depending on the transverse extension of them. Rays have a null transverse dimension and therefore can in principle describe the field with infinite resolution. Beams (tubes of flux) have a finite transverse dimension because a space discretisation is usually adopted. Gaussian beams, however, have an infinite extension, but still the field is concentrated around an axis [TaLH02]. If beams are adopted, a limit to space resolution is set. Models adopting rays are usually referred to as RT models. Models adopting beams are usually referred to as BL models or sometimes beam tracking models. In RT algorithms the position of both terminals is specified from the beginning at the generic iteration. Then, by means of geometric considerations, all the possible rays that reach the RX from the TX according to GO rules and up to a given maximum number of successive reflections/diffractions (often called *prediction order*) are found. In BL algorithms the position of only one terminal (the TX) is usually specified at the beginning. Then the space is discretised into angular sectors, each one corresponding to a beam which is launched forward in the space. Each time a beam encounters an obstacle it is reflected/diffracted according to GO rules. In BL models unfortunately, space resolution generally decreases with distance from the TX. For what was previously described, RT is suitable for accurate, point-to-point prediction while BL is more suitable for coverage prediction over large areas. Both RT and BL reproduce to a certain extent the multipath propagation process. Multidimensional phenomena related to multipath propagation (time and angle dispersion, space decorrelation, fast fading, etc.) can in principle be reproduced by ray models. However, due to limitations in environment description and/or in the modelling of the propagation process, multidimensional performances of today's models still need to be improved [Bert02].

High computation time, which depends nearly exponentially on both input database size and prediction order, represents the main drawback of ray models. Therefore a great effort has been made in trying to speed up ray models either reducing the dimension of the problem from 3D to 2D or 2.5D, or somehow limiting ray/beam tracing to a subset of dominant rays. In 2D models, propagation is assumed to take place on a horizontal plane almost parallel to ground and containing the radio terminals. This is strictly true, however, only if the radio terminals are located at the same height and if the Over Roof Top (ORT) paths are negligible. In 2.5D models instead of performing a full 3D prediction, RT is on two or more two-dimensional planes. Usually two planes are considered: the horizontal plane and the vertical plane, the latter being a plane perpendicular to the ground, containing the radio terminals and the main ORT radio paths. An interesting overview of different kinds of ray models is reported in [Bert02]. In particular an original 2.5D method called Vertical Plane Launch (VPL) is described in the paper. Since rays satisfying GO rules belong to folded vertical planes (see Figure 4.7-10), and the projection of these planes onto the ground plane results in piecewise linear lines also satisfying GO rules as rays in 2D, the basic idea is to perform a 2D RT in the horizontal plane and then to find the actual 3D paths of the rays in the unfolded vertical planes by analytical treatment. VPL is shown to perform well in field strength predictions when compared with measurements carried out in Helsinki. In particular the mean error is of a fraction of dB and the standard deviation of the error is of about 4 dB for all measurement routes. Performance is not as good as regards delay spread prediction.

A sophisticated ray tool for simultaneous prediction of both outdoor and indoor radio coverage is presented by Lostanlen *et al.* in [LoCo02a], [CoLA03] and [LoCo05]. The model is based on 2D BL in the horizontal plane combined with a multiple knife edge diffraction Deygout method to include the radial ORT path in the vertical plane [Deyg66]. The method is therefore of the 2.5D kind. The adoption of ray launching in the horizontal plane saves CPU time w.r.t. RT for predictions over areas. Moreover, the model can treat outdoor to outdoor penetration and predict indoor coverage for outdoor BSs as well. Building penetration and path loss inside buildings can be computed in two different ways. If no accurate indoor database is available, each dominant ray impinging on an outer wall of the building is then prolonged inside the building according to the following empirical-statistical formula:

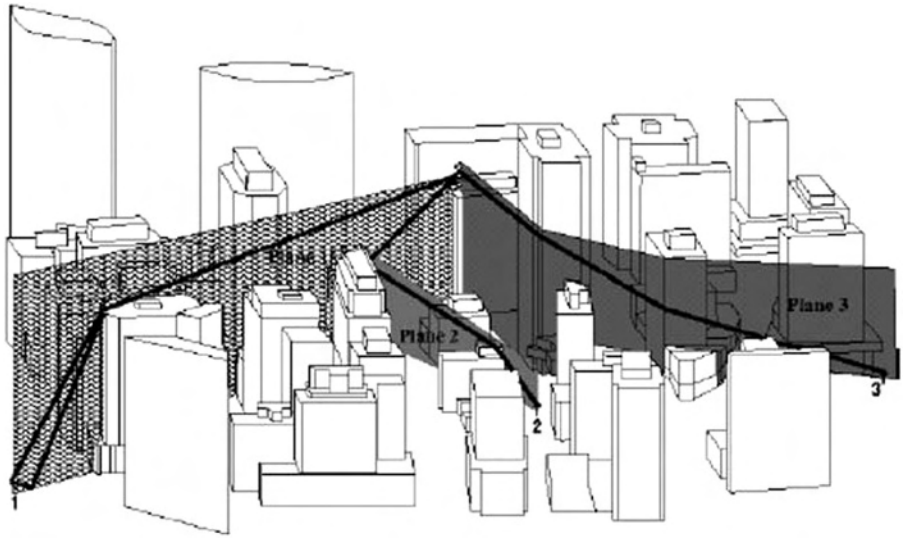


Figure 4.7-10 Example of VPL RT.

$$L_{in-building}[\mathrm{dB}] = L_{outside}[\mathrm{dB}] + L_{interface}[\mathrm{dB}] + d_{indoor}[\mathrm{m}] \times L_{linear}[\mathrm{dB/m}] \quad (4.7.1)$$

where $L_{outside}$ is the path loss calculated outside before penetration, $L_{interface}$ is the average loss due to the transmission through the facade, d_{indoor} is the path length inside the building and L_{linear} the average loss due to the in-building propagation taking into account furniture and partition attenuation. An example of outdoor-to-indoor coverage prediction derived using formula 4.7.1 is reported in Figure 4.7-11. It is evident that coverage in the fifth floor is better than in the ground floor, as can be easily experimented in the real world.

Outdoor-to-indoor propagation is especially important in public mobile radio systems (GSM, UMTS) and Digital Video Broadcasting systems, where most users are often located in indoors. In [LoCo02b], [CoLA03] and [LoCo05] also a quasi-3D RT tool is used for indoor prediction in case of outdoor and/or indoor BSs (WLAN networks). The quasi-3D RT tool, which is perfectly integrated with the 2.5D outdoor one, adopts a 2D approach in the horizontal plane (a plane parallel to each floor) plus reflections on the floor and the ceiling. Of course the indoor RT model allows the detailed

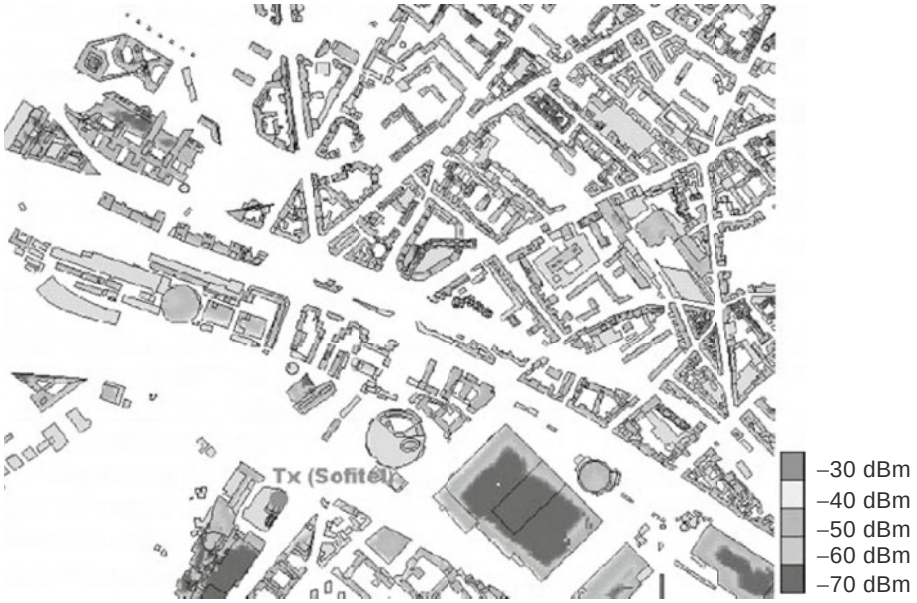


Figure 4.7-11 Examples of outdoor-to-indoor coverage.

computation of indoor multipath effects. In [LoCo05], however, it is shown that the empirical-statistical method (formula 4.7.1) outperforms 3D RT as regards outdoor-to-indoor coverage prediction. This is probably due to the great degree of environment cluttering indoors (furniture, etc.) not taken into account by the RT model, which only relies on architect Computer Aided Design (CAD) plans.

Also [JePK05] considers 3D RT for indoor propagation prediction. In this work, a full 3D RT tool, which considers multiple reflections, edge diffractions using UTD coefficients (see Subsection 4.7.2) and transmissions through walls, is used in combination with measurement data to determine the electromagnetic parameters of walls (see Subsection 4.7.3.2). This is a very important application of ray models, which thanks to their adherence to the actual physical process can be used also to model and study multipath propagation.

In [TaLH02] an interesting method based on a Gaussian BL technique is presented as an alternative to RT for 3D physical modelling of multipath propagation. The method is applied to the indoor case. By adopting the Gabor frame-based decomposition of source fields it is possible to represent radiated fields as a superposition of elementary Gaussian functions which are translated in the spatial and angular-spectral domains. Gaussian beam tracking through multiple reflections and transmissions is performed in a similar way to conventional BL methods. Fields can then be evaluated by summation of analytical terms representing transformed Gaussian beams. Diffraction is not taken into account, but due to the paraxial concentration of the field in a Gaussian beam, diffraction should have less influence with respect to the case of a plane or spherical wave. It must be noted that the parameters of a given transformed Gaussian beam have to be calculated only once for all observation points: computation time is therefore independent of the number of points where the field has to be evaluated. Simulations of amplitude-delay profiles are performed at 60 GHz, the considered indoor environment being a small laboratory area ($3.5 \times 6.5 \times 2.8$ m^3).

In order to validate the simulated results, measurements were performed in the 59–61 GHz frequency band with a network analyser. Both for simulations and measurements, the transmit and receive antennas were open ended waveguides. The measurement campaign consisted of amplitude/phase measurements at 10 to 30 different sampling points within four spots, chosen to illustrate different types of channels (Line-of-Sight, LoS, Non-Line-of-Sight, NLoS). Root Mean Square (RMS) delay spread values computed from simulated amplitude-delay profiles and from measurement results are shown in Figure 4.7-12. For clarity, only rectangles containing the values obtained for all points in a given spot are shown in the figure, together with the mean value for each spot. Not surprisingly, simulations seem to underestimate channel dispersion, but the relative difference between measured and simulated results, which is minimum for the LoS case (RX1), is coherent with previously published results in similar environments and frequency range.

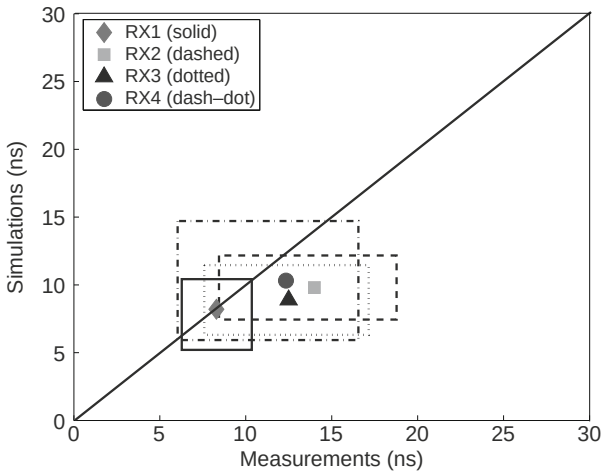


Figure 4.7-12 RMS delay spread: comparison between simulations and measurements.

4.7.3.5 Advanced techniques for ray models

As highlighted in previous subsections, ray models are generally characterised by a high computation time, especially when large input databases are needed, as for the case of macrocellular urban environments. Moreover, it is known that conventional ray models are not able to fully reproduce a correct space and angle distribution of rays, with a consequent detriment to wideband and multidimensional prediction performance. Besides the studies described in previous subsections, such as database simplification studies, fast BL models, etc., several recent studies within COST 273 specifically dealt with the development of techniques aimed at decreasing computation time and improving capabilities or performances of ray models.

Speed-up techniques

In [HoWW03] a novel method for the acceleration of 3D ray models is presented. As the database of the considered urban topology remains the same during one or more 'runs' and only the position of the TX or of the RX changes, most part of the rays remains unchanged: only the rays connecting terminals change. Therefore, a sort of 'database preprocessing' can be performed which only depends on the particular topology. In a first step building walls are divided into tiles and the edges into horizontal and vertical segments (see Figure 4.7-13). After this, the visibility conditions between these different elements

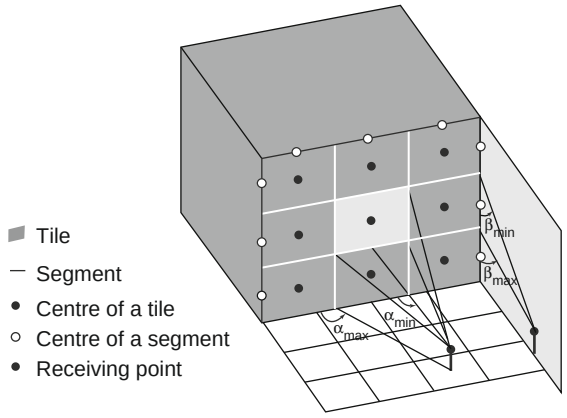


Figure 4.7-13 Tiles and segments of a wall.

(possible rays) are determined and stored in a file. The visibility relations between all tiles and segments in the database are computed in this preprocessing phase, because they are independent of the terminal locations. For this purpose all elements are represented by their centres, which leads to the discretisation of the problem of path finding, Figure 4.7-13. Based on this preprocessing the path finding can be done similar to the ray launching algorithm by recursively processing all visible elements and checking if the specific conditions for reflection and diffraction are fulfilled. The ray search is stopped if a receiving point or a given maximum number of interactions is reached. If successive runs over the same topology must be carried out, then most part of the computation is already done and saved in a file. The model is shown to perform very well in three different routes in the city of Munich, which were considered in the COST 231 blind test: the mean error is of a fraction of dB and the standard deviation of the error is of about 7 dB.

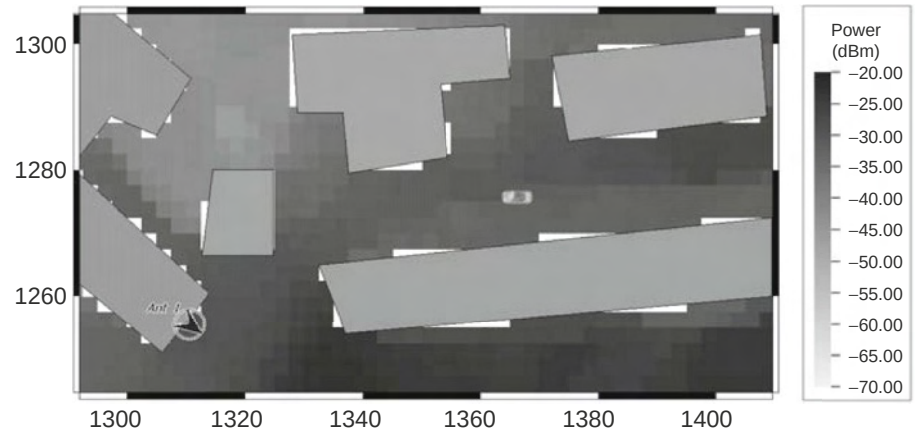
A similar performance is achieved in indoor environments. In [WZWH03] the same method as above is used for outdoor prediction, then another BL model is used for enclosed spaces with a proper interface between the two models enabling the computation of the transition from an urban scenario to an indoor scenario and vice versa, thus allowing an accurate computation of the field strength or received power inside vehicles or buildings. As a border for the interface, a polygonal cylinder including the whole indoor database is automatically determined. As an adequate way to handle the interface between the two propagation models, the Average Power Delay Profile (APDP) is used. It includes field strength values, delays and angles of incidence of the electromagnetic waves impinging on the outer walls of the vehicle or building. The reason for the need for a transition between the two propagation models is that for the indoor or vehicle objects a smaller discretisation has to be used than for the urban scenario. Prediction results in terms of field intensity have been compared with measurements: the agreement is very good, comparable to what is obtained in [HoWW03]. In Figure 4.7.14 an example of prediction of indoor coverage by an outdoor BS is shown. It is evident that the field penetrates the car mainly through the windows, therefore providing a relatively good coverage in the passenger compartment.

Diffuse scattering studies

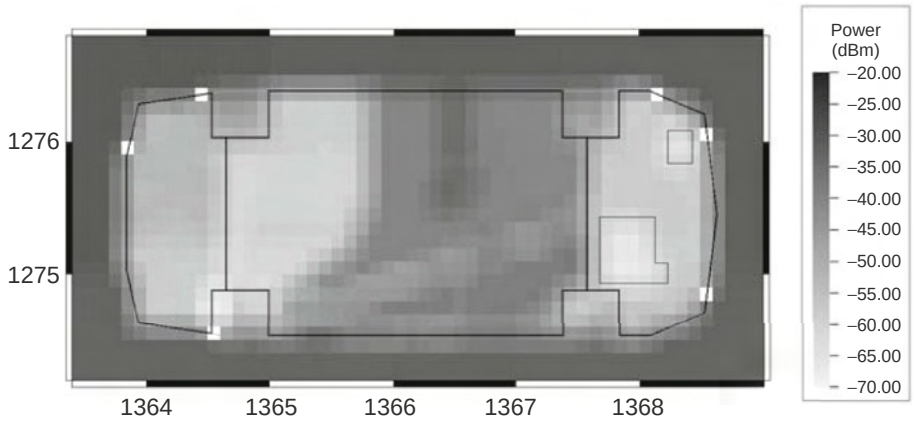
Conventional ray models only account for rays that undergo specular reflections or diffractions, but neglect diffuse scattering phenomena which can have a significant impact on propagation. As intended here, diffuse scattering refers to the signals scattered in other than the specular direction as a result of deviations (surface or volume irregularities) in a building wall from a uniform flat layer. The modelling of diffuse scattering from building walls is quite a difficult problem, since building wall irregularities cannot always be modelled as Gaussian surface roughness, as assumed in most theoretical studies. Moreover, objects *next to* (instead of *on to)* building walls such as street signs, traffic lights, advertising boards, trees, etc. can all contribute to non-GO scattering in the urban environment. A growing interest has been focused recently on the possibility to develop some sort of ray theorisation (or characterisation) of diffuse scattering phenomena. In [Bert02] an interesting study using an electromagnetic method to determine the characteristics of diffuse scattering from typical buildings facades is reported. The study demonstrated that diffuse scattering is concentrated on Keller's cones corresponding to diffractions from horizontal and vertical edges of external decorations and periodical structures such as indentations, windows, etc.

A recent experimental investigation [DEFVG05], however, showed that a single-lobe, continuous scattering pattern well describes the average scattering behaviour of typical building walls when observed from a distance. This is probably due to the fact that a considerable percentage of power penetrates into the building, interacts with furniture and inner structures, and then comes out again in a chaotic way apparently without following diffraction/reflection rules.

In [BHHT04a], the authors used a sophisticated 'synthesised array' method to determine intensity, angle and time of arrival of waves scattered from a brick building wall with windows. The wall was illuminated with a microstrip antenna, then a spatial scanning in front of the wall on a rectangular grid was carried out with the RX antenna (a microstrip antenna as well) so as to configure a 10×10 element synthesised uniform rectangular array. The output of the virtual array was then stored and processed with an ESPRIT superresolution technique in order to identify each single scattered wave and to extract the corresponding parameters. Results show that, beside the specular contribution, the wall generates



(a) Coverage in the neighbourhood (buildings in grey and base station marked with an arrow).



(b) Coverage inside the car.

Figure 4.7-14 Penetration into a car of field from a GSM BS.

a multitude of scattered waves: many weak waves are ascribed to 'brick roughness' and a few strong waves are observed coming from window metal frames.

A similar investigation is carried out in [GhTI04] with the aim of identifying the dominant scatterers in a typical urban microcellular environment. A series of measurements in a dense urban area in Yokohama city, Japan, has been carried out. The transmitter (TX) and receiver (RX) antennas were both mounted at a height of 3 metres in an LoS street scenario. The street width was 26 metres and both TX and RX were located 5.5 metres from walls on the same side of the street. Surrounding buildings had an average height of 20 metres. The measurements were accomplished during midnights with a very low traffic in the street. Employing a rotating directive antenna and a sliding-correlator at the RX the authors were able to obtain a Power Delay Profile (PDP) for every DoA. Analysis of PDPs revealed that most contributions arrive in time clusters. By comparing a precise map of the area including all present objects with the spatial distribution of received power it was observed that a significant amount of received clusters are scattered waves with DoAs corresponding to a variety of objects such as signboards, traffic signs, lampposts, traffic lights and almost any metallic object in the environment. Power considerations suggest that in a dense urban area scattering from these objects sums up to around 20% of the NLoS received power.

Recent studies have shown that diffuse scattering plays a fundamental role in determining time and angle dispersion of radio signals in the urban environment [DeBe99] [LKTH02]. Several authors have therefore tried to include diffuse scattering effects into ray models to improve multidimensional and wideband prediction performance. The first such study within COST 273 is reported in [DDGW03], where a model for diffuse scattering from surface roughness to be integrated into deterministic BL models is proposed. It is known that some portion of the energy impinging on building walls is scattered in directions other than the specular one due to the roughness of the surface. The scattered energy is generally split into the so-called *coherent* and *incoherent* components [UlMF96]. The coherent component

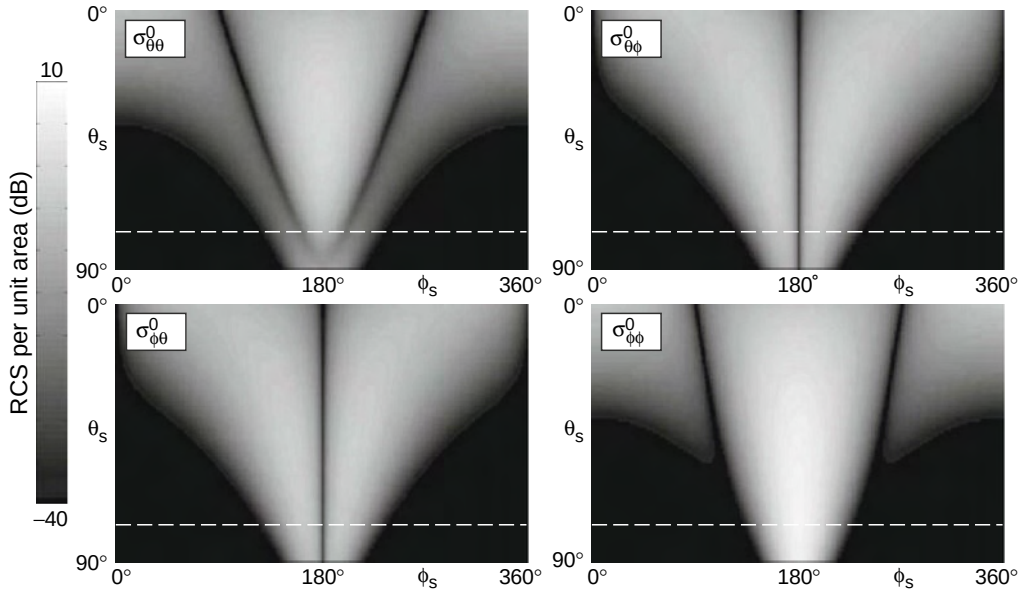
decreases with increasing surface roughness, whereas the incoherent (or diffuse) component becomes more significant. Since rough surface scattering is generally a stochastic process (with the exception of a deterministic – and therefore known – surface structure), conventionally only its mean components can be included in deterministic ray-based propagation modelling. These mean components can be determined by the Kirchhoff models with scalar and stationary phase approximation or by more complex integral equation methods. The technique proposed in [DDGW03] expands the deterministic modelling by a 'stochastic' component resulting in instantaneous realisations of the scattering process, not only its mean values. Similar to the Kirchhoff formulations, the approach is based on a tangential plane approximation, i.e. it is applicable to surfaces with gentle undulations, whose horizontal dimensions are large compared with the incident wavelength. However, in contrast to the Kirchhoff models, which are only valid for either slightly rough or very rough surfaces, the proposed stochastic scattering approach includes both the coherent and incoherent components at the same time. Using ray launching, each locally plane wavefront is actually represented by multiple discrete rays instead of only one ray. Instead of reflecting all these discrete rays at the same boundary plane, the orientation of the plane (i.e. its normal vector) and its position (i.e. its height) are varied statistically for each discrete ray and for each reflection. In that sense each discrete ray is seen as a representative of an elementary wave for the locally plane incident wavefront. The variations of the local tangential planes are directly related to the statistical behaviour of the roughness of the corresponding surface. These tangential planes are 'local' in the sense that for each discrete ray a different tangential plane is generated.

The new technique is compared to the analytical Kirchhoff model in Figure 4.7-15. A relatively rough surface with a standard deviation of the surface height $\sigma_h = 0.710$, correlation length $L = 4\lambda_0$, relative permittivity $\varepsilon_r = (10 - j3)$ and relative permeability $\mu_r = 1$, of size $Lx = Ly = 8\lambda_0$ is considered in this example. For the ray-optical stochastic scattering approach, the isotropic source is situated at $\theta_i = 60°$ and $\phi_i = 0°$, at a distance of $r_i = 2000\lambda_0$ from the plane. The rays impinge separated by $\lambda_0/10$ in the x- and y-direction on the surface, resulting in $N = 81 \times 81 = 6561$ rays per realisation. The resulting field is determined in the upper hemisphere at a distance $r_s = 2000\lambda_0$ from the centre of the plane, corresponding to approximately eight times the far-field distance, with a quantisation of $\Delta\theta = \Delta\phi = 1$ degree. A total of $M = 10^5$ realisations is taken for the ensemble averaging.

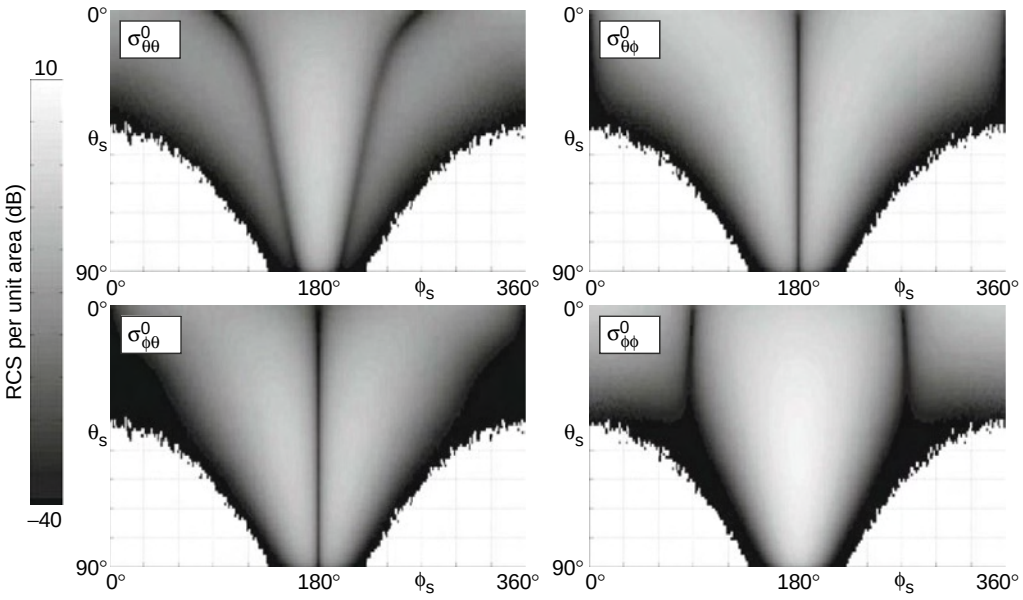
Figure 4.7-15 depicts the components of the Radar Cross Section (RCS) matrix per unit area of the diffuse (or incoherent) scattering component for the Kirchhoff model with stationary phase approximation, Figure 4.7-15(a), and for the stochastic scattering approach, Figure 4.7-15(b). The broken line in Figure 4.7-15(a) indicates the limit of the validity region of the Kirchhoff model. The absence of colour in the lower corners in Figure 4.7-15(b), indicates that no ray reaches these regions in the stochastic scattering approach. The absolute value in the specular direction of, e.g. $\sigma^0_{\phi\phi}$ is in both cases 6.3 dB. The coherent scattering component is negligible for this type of surface (more than 50 dB below the incoherent component). Both methods show the same mean scattering behaviour: a more distinct main-lobe for the co-polar $\phi\phi$ *component* compared to the $\theta\theta$ *component*, a maximum of the scattered energy in a region near the direction of the specular direction, and a minimum of the cross-polar components in the specular f-direction. The comparison of coherent scattering from less rough surfaces shows a good agreement between the Kirchhoff model and the stochastic scattering approach as well. By applying this stochastic scattering approach, the inclusion of random (but directive) surface scattering into ray-optical propagation modelling becomes possible. The purely deterministic GO modelling is expanded by a 'stochastic' component, resulting in varying prediction results; thus allowing for the first time to account for non-deterministic scattering in ray-optical wave propagation modelling.

In [DEGDA04] and [DEESGK02a] an approach for the integration of the diffuse scattering contribution from both surface and volume inhomogeneities into RT models is proposed. The so-called Effective Roughness (ER) approach first proposed in [DeBe99] and [Degl01] is implemented in a 3D RT model and results are compared with measurements. According to the ER approach diffuse scattering, although mainly due to surface and volume irregularities (windows, balconies, indentations, decorative elements, rain-pipes, internal reinforcements, power lines, heating pipes, etc.), is assumed originating from surface roughness scattering. An *effective roughness* is associated to each wall in order to take into account not only real surface roughness but also the mean effects of the above-mentioned elements. A Lambertian scattering pattern is ascribed to each wall (if the wall is far enough from TX and RX) or to each surface element (if the wall is close). The scattering contribution of each wall is computed directly from wall distance and orientation with respect to the TX and the RX using simple, analytic formulas derived from power balance considerations, which depend on only one parameter, the scattering parameter S. S is defined as the ratio between the amplitude of the scattered field and the amplitude of the incident field in the vicinity of the surface. In case of a far wall, the square of the total scattered field amplitude is:

$$E_s^2 = K_0^2 S^2 \frac{A\cos\theta_i\cos\theta_s}{\pi}\frac{1}{r_i^2 r_s^2}; \quad K_0 = \sqrt{60 G_t P_t}, \tag{4.7.2}$$



(a) Computed with the Kirchhoff method.



(b) Computed with the ray-optical stochastic approach.

Figure 4.7-15 RCS matrix elements per unit area ($\sigma^0_{\theta\theta}$, $\sigma^0_{\theta\phi}$, $\sigma^0_{\phi\theta}$, $\sigma^0_{\phi\phi}$) of a square rough surface. Direction of illumination: ($\theta_i = 60°$, $\phi_i = 0°$).

where G_t and P_t are the gain and the input power of the TX antenna, respectively, and A is the area of the wall. Angles and distances in (4.7.2) correspond to that depicted in Figure 4.7-16. In order for the overall power balance to be satisfied, specular reflection and diffraction are consequently attenuated according to the Rayleigh factor R. Realistic S and R values for typical buildings have been found to be $S = 0.4$ and $R = 0.6$ [DeBe99]. For simplicity, only first-order scattering is considered. Results reported in [DEGDA04] show comparisons of measured and simulated power delay profiles relative to

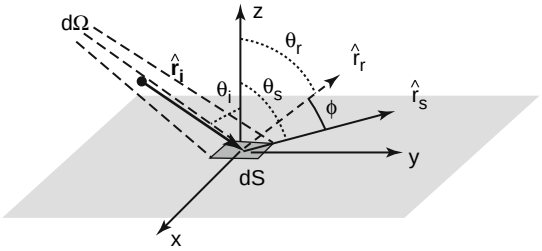


Figure 4.7-16 View of the generic surface element with incidence (θ_i) and scattering (θ_s) angles.

a macrocellular scenario in the city of Stockholm. The diffuse component is shown to be of major importance, especially for wideband assessments in macrocells. In Figure 4.7-17, for example, the agreement is quite good both with $S = 0.4$ and with $S = 0.6$. On the contrary, results without scattering obtained with an ordinary 2D RT model are unacceptable. Note that 2D RT results do not include the first peak, which must of course be computed using an ORT model. Therefore, using a traditional *2D RT+ ORT* model the overall received power would indeed be predicted with a good accuracy while wideband parameters such as *DS* would come out heavily underestimated due to the strong underestimation of the 'tail' of the PDP. Similar results but relative to a more complete characterisation in terms of both delay spread and angle spread of the multipath channel in the city of Helsinki are reported in [DEESGK02a] (see Section 4.7.5).

An interesting study on the role of scattering in ORT propagation is reported in [DeFG03]. In dense urban environment, where the average building height is comparable to or even greater than street width, multiscreen diffraction only partially accounts for the overall ORT propagation mechanism. In particular, the last part of the ORT path, the so-called Roof-to-Street (RtS) part, is a complex phenomenon also involving multiple reflections and scattering on the buildings surrounding the MT. Several authors proposed 2D RT in the vertical plane with multiple reflection/diffractions to model RtS propagation. Reflections and diffractions alone, however, are quite inefficient for RtS propagation in dense urban environment since the former requires a large number of bounces and the latter involves deep-shadow diffraction, which is very weak. According to preliminary studies reported in [DeFG03], where simulations of the RtS mechanism in typical cases are carried out using RT including ER scattering in the vertical plane, diffuse scattering from building tops and upper parts, which are often directly illuminated by the BS, plays a key role in determining the RtS contribution.

UWB propagation modelling by RT

UWB is a very promising technique for high-data rate, short range transmission and is claimed to permit very low power consumption and therefore very high battery duration in portable terminals. In [LoGC05] a 3D RT model is applied to the modelling of UWB propagation in an indoor environment. The application of ray models to UWB propagation prediction is quite straightforward. Since the geometry of the multipath pattern does not depend on frequency, the rays are generated by the RT algorithm in a similar manner to the narrowband case. On the contrary, the amplitude and the phase/delay of each ray do depend on frequency because of the dependence on frequency of reflection and diffraction coefficients and of the emission/reception characteristics of the antennas. A UWB time-domain-pulse emitted by the TX is therefore not only replicated due to the multipath but also distorted due to the above-mentioned characteristics. The effect of multiple bounces inside the dielectric slab on a UWB pulse transmitted through a wall is also shown in the paper. Since the antenna patterns must be considered frequency dependent, a time-domain antenna diagram must be input to the RT simulator. An example of the effect on a transmitted pulse of the overall indoor propagation process is shown in Figure 4.7-18.

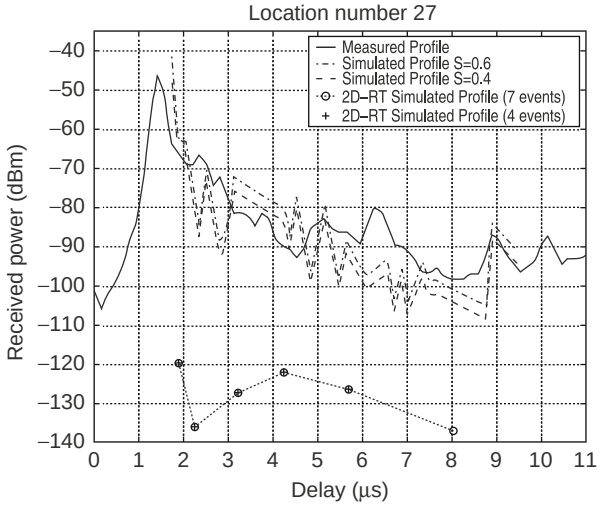


Figure 4.7-17 Measured and simulated PDP with and without scattering.

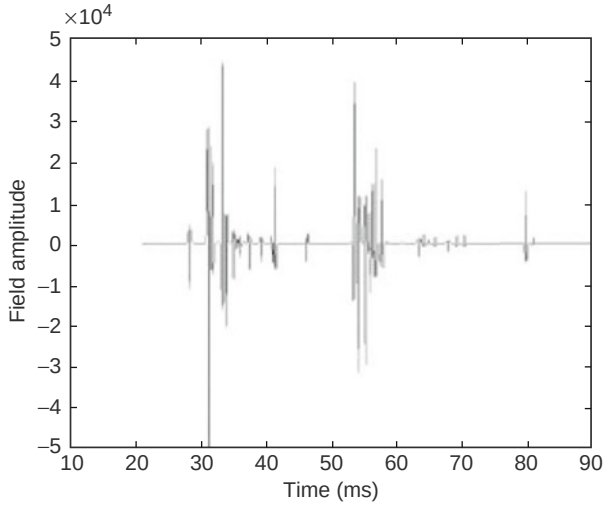


Figure 4.7-18 Received UWB impulse with TX and RX in NLoS on the same floor in a simple office environment.

4.7.3.6 Applications of ray models

Since ray models can potentially reproduce the multipath propagation process, the range of possible applications is virtually unlimited. Power coverage predictions can of course be achieved with ray models. However, simpler empirical-statistical models such as Hata-like models, multiwall models for indoors, etc. can provide a nearly similar level of accuracy in most cases with a lower computing time. Ray models are therefore worthwhile if a higher insight into the multipath process is needed.

This is the case of fading statistics assessments, wideband assessments, multidimensional (space, time, frequency, Doppler frequency) channel characterisation, and any other problem involving these aspects. In [Bert02] and [ChBe01] a VPL RT model is used to perform radio channel characterisation. Channel statistics are needed for advanced system design, but measurements are costly. Moreover, measurements carried out in one environment may not apply to a different environment. The use of ray models is a relatively inexpensive substitute for measurements, with the advantage of allowing the simulation of a large number of cases in a very short time. In [Bert02] it is shown how it is possible to derive channel statistics on the base of the output of a large number of RT runs by means of a Monte Carlo approach. In particular, DS, AS, finger life statistics in rake receivers and other statistics are obtained and reported vs. environment characteristics, building height statistics, etc.

In [TLVD01] fading statistics are obtained by RT simulations. In particular, starting from the multipath information at one single point, it is possible to estimate the signal statistics in a local area of that point, therefore an efficient site-specific channel model might be feasible. Using the RT technique, the Doppler power spectral density $S(f)$ for a receiver which is assumed to be moving in a certain direction with a given velocity can be computed from direction of arrival and intensity of each path. Moreover, making use of the phase of each wave ϕ_k, the complex function $R(f)$, where $|R(f)|^2 = S(f)$, can be derived and the complex envelope of the received signal in the time domain $r(t)$ can be obtained by simply performing the inverse Fourier transform of $R(f)$. This process is obviously much faster than the simulation of the received signal in a great number of locations (denoted as method 1 in this work). Finally, once the complex envelope of the received signal in the time domain $r(t)$ is known, the spatial distribution of the fades is obtained by considering that time and space are obviously related by the velocity (see formula (4.7.3)). The fast fading statistics in a local area around the point where the Doppler spectrum has been calculated are obtained by suitably sampling the curve $\tilde{r}(l)$

$$\tilde{r}(l = vt) = |r(t)| = \sqrt{r_I^2(t) + r_Q^2(t)} \qquad (4.7.3)$$

It is important to note that the velocity of the mobile is used here as a parameter whose actual magnitude is not relevant, so that it can be considered as a differential velocity. The spatial distribution of fades depends only on how the power density is distributed along the Doppler frequencies, but not on what the maximum Doppler frequency is.

A wide range of comparisons between measurements and RT simulations has been carried out [TLVD01]. Part of the measurement campaign was carried out in the 5th floor of the Communications Engineering Department of the University of Cantabria, a very complex environment. Three RX trajectories, T1, T2 and T3, are analysed here. Both the data obtained from the measurements and those resulting from the simulations were processed in order to obtain the Rician probability distribution function that best fitted the curve in each case, according to the criterion of the minimum root mean square error. Figure 4.7-19 shows the Rice K-factor of the local fast fading estimated for these three trajectories. Figure 4.7-20 presents the normalised level crossing rate for the trajectory T1, this case being representative of all the others. It can be observed that the results obtained, from both the measurements and the simulations, closely approximate the theoretical ones since this propagation environment is close to the case of isotropic scattering.

A different approach, which can be used to derive local fading statistics (including the expected mean level), is presented in [NaCB04]. For network planning purposes, it is of interest to estimate the expected local mean level. Due to multipath fading (often referred to as Rayleigh fading), it is necessary to average large blocks ($10\lambda \times 10\lambda$) of hundreds of points separated by a fraction of wavelength (e.g. $\lambda/4$) to estimate the local mean level, and this can require enormous computational

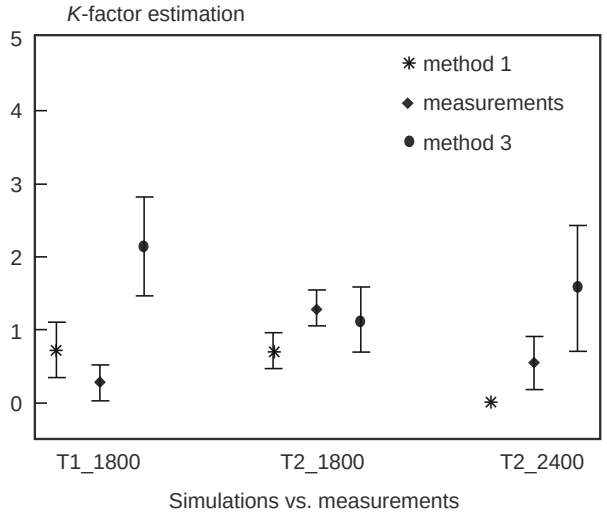


Figure 4.7-19 Measured and simulated Rice K-factor for the three routes.

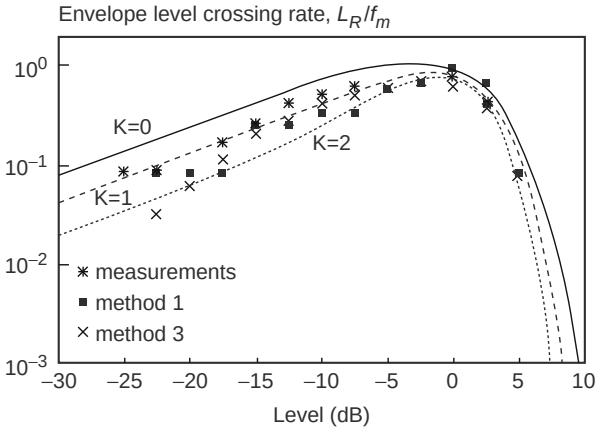


Figure 4.7-20 Measured and simulated Rice K-factor for the three routes.

resources. An alternative approach to estimating the expected local mean level, based on artificially modulating the wall surfaces to cover all the possible signal phases, is the proposed wall imperfection model. The proposed model is implemented by randomly varying the position of each wall using a uniformly distributed random variable over a number of 3D RT runs in a simple indoor environment. The standard deviation of the random variable can be varied to study different degrees of wall displacement. The uniform distribution is selected to calculate the local mean field strength by averaging out all phases with equal weighting. A similar methodology can be used to study imperfections of wall material properties by adding a random variable with an arbitrary distribution (uniform or Gaussian) to the electromagnetic wall property values to account for uncertainties in these values. This proposed approach in estimating the average local value of the signal strength has the advantage of decreasing the number of computer runs compared to the approach of averaging blocks of points with high resolution (λ/4) grids. The wall imperfection model was tested for several environments. For example, the model is applied to estimate the signal level at a central RX point (15 m, 15 m)in an arbitrary indoor environment of 30 m × 30 m. The procedure is described as follows. During each computer run, all the walls are moved simultaneously and randomly to give uncorrelated wall movements and the phasor sum of all the rays arriving at the receiver is obtained. An optional additional random variable with a 10% standard deviation (selected arbitrarily) is added to the values of the electromagnetic properties of the walls (permittivity and conductivity) to account for uncertainties in these values. The values of n computer runs are then averaged to obtain an estimate of the local mean field strength at the RX point. The estimated level is compared with the value obtained by averaging a block of points around the receiver separated by less than λ/4. The obtained averaged field value is compared for different computer runs (for the wall imperfection model) and different number of points forming the averaging block (for the spatial sampling approach). The results are summarised in Table 4.7.3. The results from averaging blocks of points following the spatial sampling approach suggest that averaging 400 points (5λ × 5λ) seems to be enough for indoor environments to reach a stable average value after minimising the effect of small signal variations. This contrasted with the traditional approach used in outdoor environments of averaging larger blocks of (10λ × 10λ) or (20λ × 20λ). The average field level obtained from the wall imperfection model seems to have reached its final value after 50 to 100 runs, therefore with a reduction of 87.5% in the number of computer runs.

In [FuMa03] a full-3D RT tool is used to perform a wideband characterisation at 2 GHz in a typical urban macrocellular scenario, Figure 4.7-21. The model considers multiple reflections, multiple diffractions and combinations of both. Here the maximum reflection order is set to six and the maximum diffraction order to two. The simulation environment is a 719 × 539 m section of the city of Karlsruhe and 'BS' marks the place of the BS (transmitter), which is located on a 12.5 m high building on a pole 13 m above the roof. The receiver

Table 4.7-3 Comparison of the average local field value around the receiver obtained from the wall imperfection model and the spatial sampling approach for different computer runs.

(a) Wall imperfection model		(b) Spatial sampling approach	
Number of computer runs	**Average field value [dB]**	**Number of runs (separation: λ/4)**	**Average field value [dB]**
10	−18.92	100 (2.5λ × 2.5λ)	−20.52
50	−19.39	400 (5λ × 5λ)	−19.32
100	−19.34	1600 (10λ × 10λ)	−19.30
200	−19.22	6400 (20λ × 20λ)	−19.34
300	−19.45		

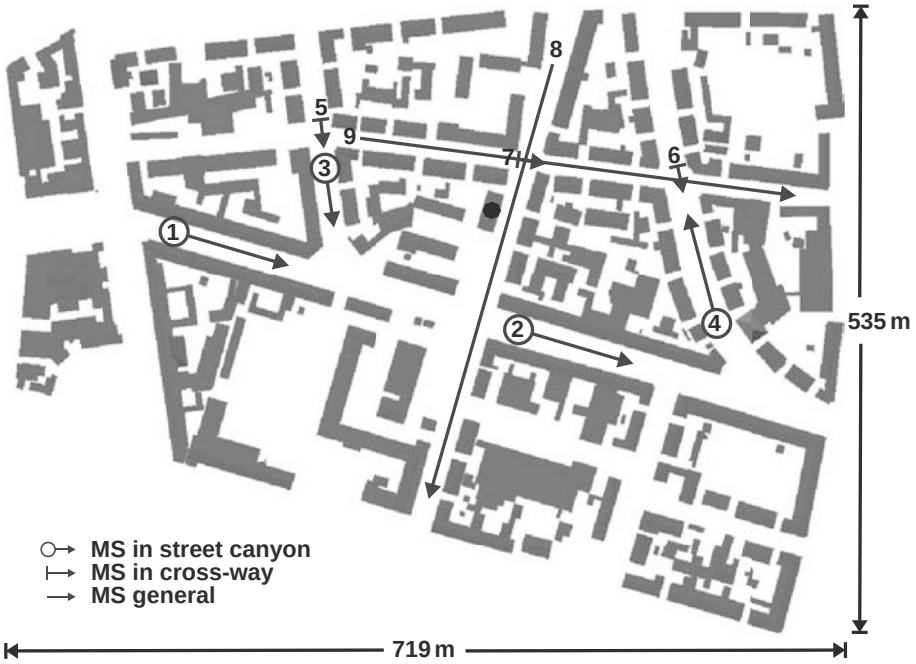


Figure 4.7-21 Top view of the simulation environment.

(mobile station – MT) has an antenna height of 1.7 m above ground. A λ/2 dipole is used as both transmitting and receiving antenna. The channel impulse responses are collected in nine different simulation routes, separated in three different propagation classes: [RX1]–[RX4] capture street canyons, [RX5]–[RX7] intersection and [RX8]–[RX9] capture all propagation classes. Considerations regarding the behaviour of the multipath pattern in different route classes show a good agreement with that-reported in [LKTH02]. Table 4.7-4 summarises the statistical parameters for all simulation routes. Comparing scenario I (street canyon) with scenario II (intersection), one can see that at intersections time and angle spreads are higher than in the latter. Probably, at intersections waves are not influenced by street guiding effect and thus energy is spread over a much larger area in the azimuth plane. Comparing delay spread and angular spread a high correlation can be observed. Typically correlation is higher at the BS than at the MT. The obtained parameters can be used for the parametrisation of spatial channel models.

In [CoWS04] a 2.5D RT model [RiWG97] is used as a propagation engine in a UMTS simulation program. The environment is a dense urban area in central Paris. The same UMTS simulation runs have also been done using traditional Hata-like empirical propagation models such as the well-known and tested COST-231-Hata model. Field prediction and system simulation outputs obtained with the two propagation models are then compared and discussed. In Figure 4.7-22, for example, field strength coverage with multiple BSs is shown as predicted by the Hata-like model (a) and by the RT model (b). It is evident that the empirical model, although taking into account the pattern of the antennas, is not able to reproduce the interaction of the propagating field with the urban layout, which is on the contrary evident on (b). Of course, such a huge difference reflects on system performance assessments as well. It is shown in the paper that a UMTS system simulation with the empirical propagation engine leads to less than 1% rejection rate R_{jR}, with

$$R_{jR} = (number\ of\ rejected\ calls / total\ number\ of\ calls)$$

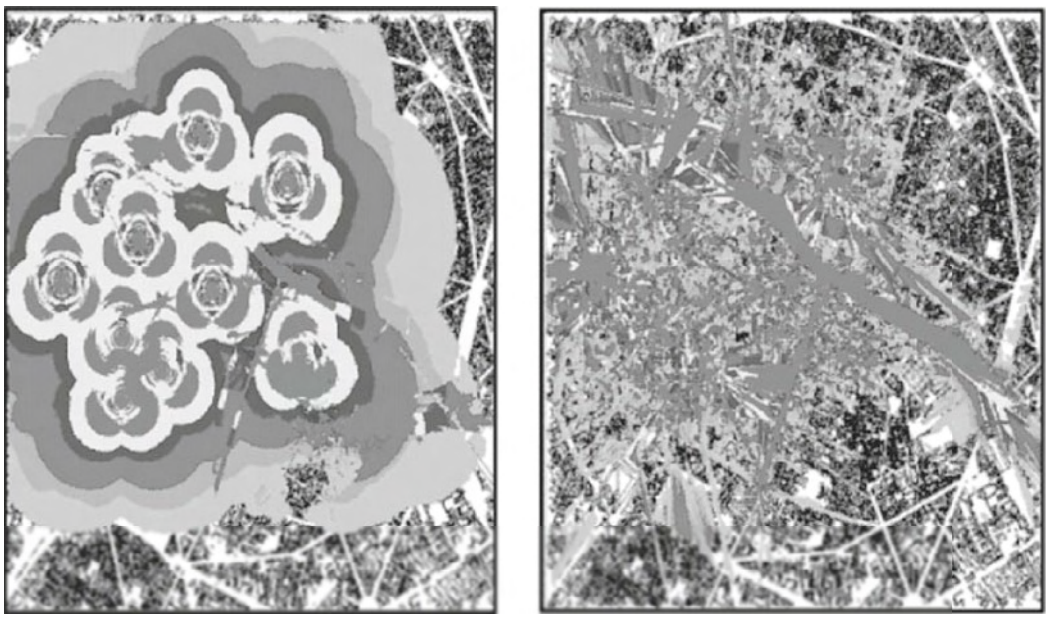
while the same simulation with RT leads to $R_{jR} = 20\%$.

A group of researchers from the University of Stuttgart and from AWE Communications, Germany, have worked on advanced localisation methods using ray models [WHZL02], [ZBLL04]. In particular, the ray-tracing model described in [HoWW03] is used to implement the so-called 'database correlation method' for localisation in an urban environment. Field prediction results obtained with RT in a given service environment are used to define a look-up table for every BS. By evaluating the measured path losses between the MT and the BSs and by correlating these losses with the entries of the look-up table, the determination of the location of mobile terminals is obtained. In order to show the achievable performance in real environments, measurements in a GSM network are used for the localisation. The measurements represent a mobile which receives simultaneously four BSs.

Table 4.7-4 Summary of delay spread, azimuth spread and correlation coefficients between delay and azimuth spread for routes [RX1]–[RX9].

Route RXN	Scenario	Delay spread (ns)	BS azimuth spread [deg]	MT azimuth spread [deg]	BS corr. coeff.	MT corr. coeff.
1	MT in street canyon	29.3	28.5	3.0	0.23	0.80
2		160.9	18.5	14.7	0.33	0.99
3		75.0	52.2	12.9	0.09	0.65
4		72.6	56.8	13.8	0.57	0.58
Mean value		84.3	39.0	11.1	0.30	0.76
5	MT in intersection	303.5	48.5	31.7	0.78	0.59
6		106.5	55.6	12.7	0.73	0.90
7		32.6	20.4	7.9	0.08	0.61
Mean value		147.5	41.5	17.4	0.53	0.70
8	MT in street canyon and in intersection	72.7	29.4	16.1	0.74	0.84
9		112.3	38.4	14.9	0.40	0.73
Mean value		92.5	33.9	15.5	0.55	0.79
Mean value of all routes (distance = 25–50 m)		108.1	38.1	14.7	0.46	0.75

The received power values lead to a pattern which is compared to the database entries to determine the location of the MT. Figure 4.7-23 shows the achieved results for the test scenario. The initial effort for the location technique (i.e. the determination of the pixel matrices and the computation of the entries of the look-up table) is higher than the effort for conventional approaches. But the achievable localisation accuracy is significantly better



(a) COST 231-Hata model. (b) 2.5 RT model.

Figure 4.7-22 Field prediction over a section of central Paris.

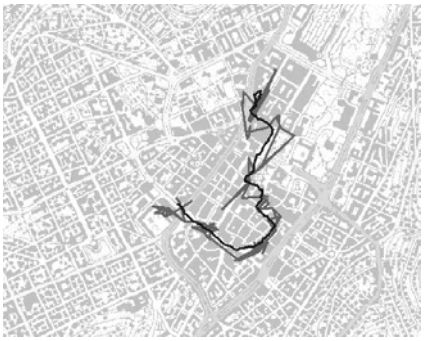


Figure 4.7-23 Real (black) and estimated (grey) route of a walk through the city-centre of Stuttgart, depicted area: 2.5 km × 2 km.

if really accurate prediction models like 3D RT or other accurate models (like the dominant path approach, see Subsection 4.7.3.7) are used.

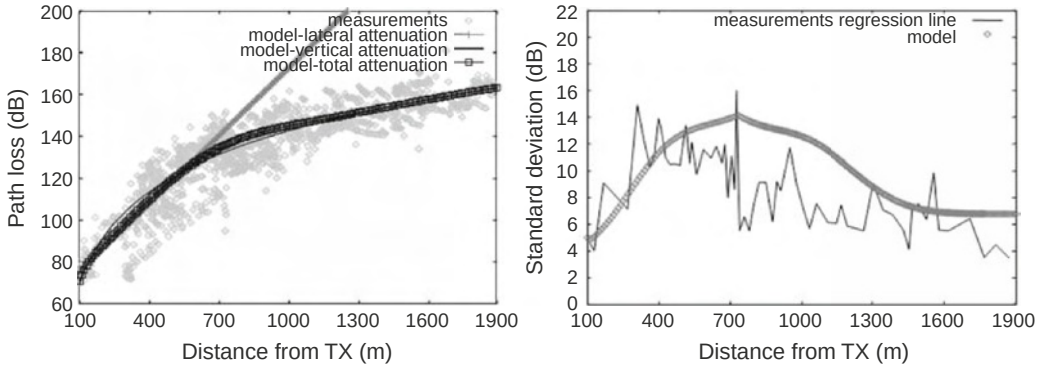
Ray models can be of great help in the optimisation phase of an existing mobile radio system. In [LBLC02], for example, the authors tried to understand the reason of a coverage gap in a 400 MHz MPT 1327 trunking system by using a 3D RT tool (EPICS). A full 3D representation of the environment, including terrain height, was adopted and the study was carried out by comparing measurements and RT simulations. The study confirmed that bad coverage was due to the effect of the buildings and suggested possible solutions to the problem.

4.7.3.7 Simplified or hybrid models

Simplified or hybrid models may represent a very attractive compromise between the complexity of ray models and the oversimplification of Hata-like models. The term 'simplified' refers to the process of progressive simplification which usually leads to the development of a simplified model starting from a complete ray model or from a set of measurement data. The required input database is often simplified too, with a consequent acquisition and handling cost reduction. The term hybrid refers to the fact that, while simplifying, statistical elements are necessarily introduced in the deterministic approach. Unfortunately, the simplification of an intrinsically complex propagation process can indeed yield simple models, but at the expense of a proliferation of arbitrary parameters, which often need to be determined using measurements or RT simulations. It is interesting to notice that some simplified models for indoor environments [LoCo05], [WWWW04] often yield better accuracy in field strength predictions than sophisticated 3D RT tools. This is probably due to the inherent calibration with measurements present in simplified models which permits them to take into account the effect of environment cluttering (furniture, etc.) as opposed to conventional ray models which rely only on architect CAD plans. In principle, a good simplified model could guarantee a good accuracy with a drastic reduction in computation time with respect to conventional ray models, although the output is often limited to field strength and fading statistics. Therefore, these models are suitable for simple field predictions involving a large number of trial runs, such as those required for mobile radio network planning and deployment assessments.

A simplified model for field prediction in the urban environment is described in [BCFF02]. The model is derived from the observation that propagation in the urban environment may occur owing to different mechanisms, namely LoS propagation when there is a direct path between the two terminals, *lateral* NLoS propagation along the streets and around the building corners, and *vertical* NLoS over the building rooftops. For LoS propagation the authors resort to the well-known two-ray propagation model (including the direct ray and the ground reflected ray), while for the two NLoS components they developed new models which benefit from a statistical characterisation of the urban layout. The overall path-loss model is then defined as a weighed sum of these components, and the model yields estimates of the field fluctuations as well. The obstruction probability, which is defined as the probability that a link between a transmitter and a receiver is obstructed, defines the boundary between the LoS and NLoS regions. Then, the contributions to the field strength propagating in the two planes (lateral and vertical) are assumed independent and each of them can be modelled with a Gaussian random variable with known mean and standard deviation. With this hypothesis the field strength and its standard deviation vs. distance result from simple formulas involving some environment parameters (q, μ_{LAT} and μ_{VERT}) which depend on synthetic topological characteristics of the considered area (e.g. mean building height, mean street width, etc.) and, as usual, are expressed in dB. The parameters' values must be extracted from the city map and from measurements or RT simulations. A comparison in terms of path loss vs. distance between measurements performed in Munich at 947 MHz (COST 231 blind test) and model prediction is shown in Figure 4.7-24(a): the agreement is quite good and we can observe that the vertical and lateral contributions, as expected, equal each other at a distance of about 650 m. Figure 4.7-24(b) reports the received field standard deviation as a function of the distance. From the figures we may conclude that the developed model is able to reproduce the experimental values of mean path loss and standard deviation with good accuracy once the values of the involved parameters are tuned to the actual environment. It is worth highlighting that the model does not require a detailed building database.

In [WWWW04] a new semi-deterministic model for field prediction in both the indoor and outdoor environment is proposed. The model is aimed at taking into account only the dominant paths in each scenario and



(a) Measured and simulated path loss.

(b) Standard deviation of the field strength (right) vs. distance.

Figure 4.7-24 Measured and simulated path loss and standard deviation of the field strength vs. distance.

reducing dependency on the accuracy of the input database with respect to complete ray models. The model requires a simple calibration with measurements or reference data. The model itself can be subdivided into two steps.

(a) Determination of the dominant paths

Determining the dominant paths is the most critical task. For indoor scenarios the algorithm is explained more in detail in [WLGB97] and [WoLa98] and a comparison with RT is shown in Figure 4.7-25. The model is able to determine the 'less obstructed path' which usually transfers most of the power to the RX. The same principle can also be used for urban scenarios. Figure 4.7-26 shows an example for urban scenarios (Hong Kong, incl. topography). The dominant paths describe the main direction of propagation (the corridors or rooms passed (indoor)), i.e. the path of the main energy flow is determined. No individual reflection or diffraction coefficients need to be computed. These interactions are included in the 'parameter' of the paths (e.g. waveguiding factor, depending on the reflection loss and distance of walls in the scenario). The paths can either be determined in a horizontal plane (2D mode) in 2.5D (urban case, Figure 4.7-26) or 3D (indoor). So in urban scenarios the wave can also propagate over the rooftops as well as in the street canyons.

(b) Prediction of the path loss along the paths

The (empirical) prediction of the path loss along the paths depends on path length d (path loss proportional to $n \times \log(d)$), on the number of interactions, optionally also on the angles in the path at the interaction points and on the model-specific parameters (waveguiding effect, explained above). These

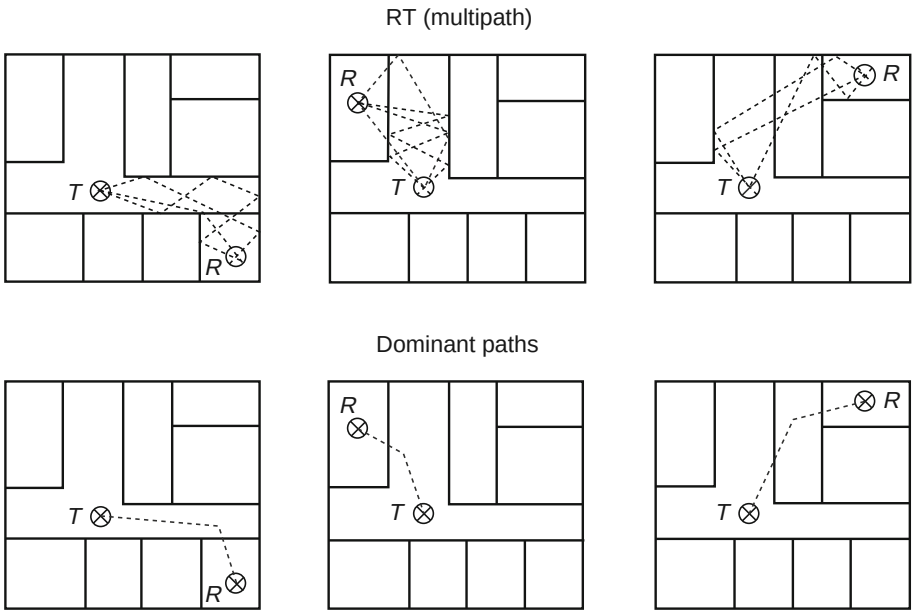


Figure 4.7-25 Example of dominant paths tracing as opposed to RT in an indoor scenario.

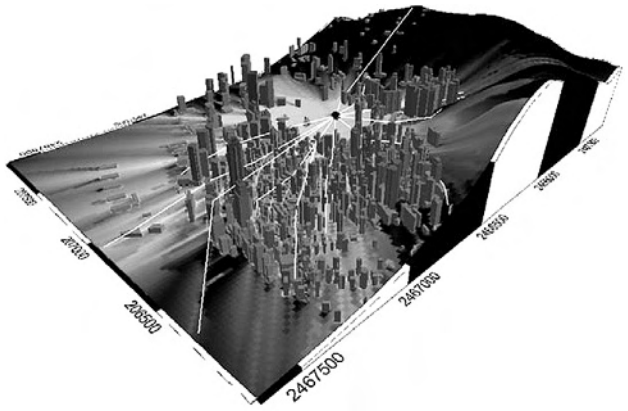


Figure 4.7-26 Example of dominant paths tracing in an outdoor scenario (Hong Kong city).

parameters can automatically be calibrated with measurements. The new model has been tested in indoor and outdoor environments by comparing field strength predictions with measurements. Once calibration is done, the attainable accuracy is very good, comparable with RT prediction (standard deviation of the error of about 4 dB indoors and 6–8 dB outdoors) with a very low computation time. The computation time for the Munich COST 231 blind test scenario (2000 buildings, 10 m resolution, 8.7 km^2 area) was below 30 seconds on a 2 GHz PC.

A formulation of path loss vs. distance for urban microcells into simple empirical statistical models is reported in [ZVRK04]. On the base of a 5.3 GHz MIMO measurement campaign carried out in central Helsinki with the BS located in the middle of street crossings at a height of about 10 m, two path-loss formulas were derived by measurement fitting using the least-square method for both LoS and NLoS situations. The BS was equipped with a uniform linear array of four dual polarised elements. Measurements from the MT were performed over rectilinear routes and fast fading was averaged out from measurements using a sliding window of about 80λ, and thereafter the powers over the MIMO subchannels were averaged in each MT location. The LoS formula is a simple Hata-like formula:

$$PL(d_1) = 40.3 + 23.4 * \mathrm{Log}(d_1)[\mathrm{dB}];$$
$$\sigma_{LoS} = 2.6[\mathrm{dB}]$$

where d_1 is LoS distance and σ_{LoS} is the Path Loss (PL) standard deviation. It is interesting to note that the PL exponent (=2.3) is greater than in free space (=2). The NLoS formula is:

$$PL(d_1, d_2) = PL(d_1) + 10_{n_j} * \mathrm{Log}(|d_2| - W/2)[\mathrm{dB}];$$
$$\sigma_{LoS} = 3.1[\mathrm{dB}]$$

where d_1 is the LoS distance from BS to the closest street intersection, d_2 is the distance from the intersection to the MT, n_j is a randomly drawn path-loss exponent with mean = 2.33 and standard deviation = 0.3 and W is the main street (where the BS is located) width.

In [HeKu05] several different deterministic simplified models are compared in a UMTS urban macrocellular environment with ultra-high sites, i.e. with BSs mounted at a height of at least 100 m. Namely, the models under comparison are COST-WI, flat edge, Maciel-Xia-Bertoni and a reference free-space formula. Since the elevation angle of the LoS path can be quite high, Maciel-Xia-Bertoni shows large prediction errors at small distances, where COST-WI performs well. A hybrid model is proposed which combines Maciel-Xia-Bertoni and COST-WI, switching from the former to the latter where the elevation angle exceeds 0.15 degrees.

The adoption of simplified models is almost mandatory when general results must be derived which are not specific to a particular topology. In [Fusc04] a simplified Hata-like model is used to study innovative coverage solutions for DVB systems. The capability of providing high quality reception by handheld terminals is a very important target in digital video broadcasting. Unfortunately, preliminary assessments and field trials indicate that this target cannot be satisfactory achieved by conventional network planning, based on the reuse of existing transmission sites (in order to limit the economic effort in the initial deployment of new digital broadcasting networks). Hence new deployment strategies must be investigated. In particular this work shows the possibility of strongly improving coverage probability in urban environment by adding some urban and suburban gap filler to conventional DVB network deployment. This is possible exploiting the flexibility of the single frequency, OFDM transmission technique. Results show that, with a conventional planning adopting extra-urban transmitters only, the minimum power at the TX stations should reach up to 95 kW to achieve a coverage probability $PC \geq 0.95$ all over a circular urban area with a radius of 5 kilometres. By adopting a new, more diffused coverage solution composed of 24 urban TXs (irradiating 8 W each), 10 suburban TXs (30 W) and 3 extra-urban TXs (5.845 kW) the same coverage probability can be obtained saving nearly 90% of the total Transmitter (TX) power. The resulting coverage map is shown in Figure 4.7-27.

4.7.4 Channel measurements and parameter estimation

4.7.4.1 Measurement scenarios

The purpose of measurements and their analysis is typically twofold. In the first place, propagation measurements are still needed to gain deeper understanding on the

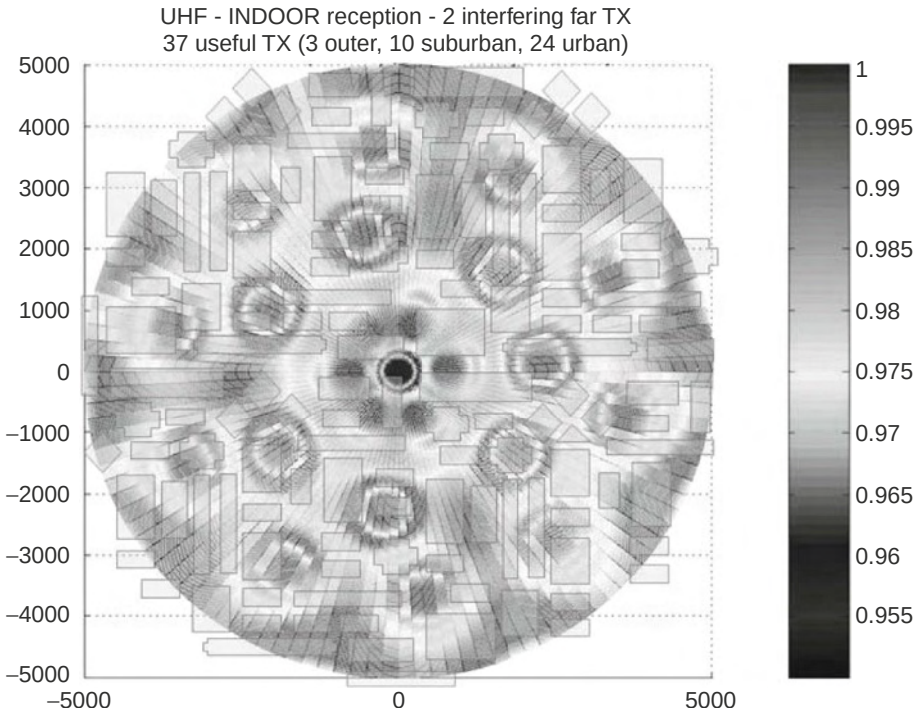


Figure 4.7-27 Coverage map with the new deployment over a circular urban area with a radius of 5 km.

very complicated combination of propagation phenomena and their mutual significance in mobile radio systems. Due to the progress of transmission technologies, more dimensions of the channels must be taken into account in the measurement. In comparison with COST 259 [Corr01], double-directional channels are extensively focused for the MIMO systems within COST 273. As a very important new aspect, the polarisation characteristics of the channel have to be known for polarisation diversity or MIMO systems. For all these aspects, one should obtain statistically adequate information in all relevant environments to be able to identify the most significant features of the propagation channel to first understand what is happening in the propagation-physics point of view and further to be able to create realistic channel models for system studies.

On the other hand, there are lots of measurements related to the systems aspects of the current and future mobile radio systems. In a radio system, the radios have to deal with several simultaneous connections between BSs (or access points) and mobiles. Thus one should have information not only on a single link but also on several radio connections and their mutual relationships. Networks are probably the ultimate example of the complexity of the propagation scenarios in the future mobile radio systems.

4.7.4.2 Channel sounding techniques

Several new architectures of the channel sounders have been proposed for the double-directional channel sounding. This subsection provides an overview of the radio channel sounding hardware architectures including array antennas for directional measurements.

Channel sounder architectures

A channel sounder consists of a TX to repeatedly transmit a fixed wideband waveform and an RX connected to a data recorder to record a waveform distorted due to the propagation channel. For short range measurements, a Vector Network Analyser (VNA) can be used instead of a dedicated sounder. Alternatively, two VNAs can be used as TX and RX respectively [MGPCBRJL04]. A cost-effective approach for SISO measurements is the use of a spectrum analyser as RX [PLVP02]. The phase information can be recovered from the amplitude spectrum by using Hilbert transform.

Various kinds of wideband waveforms are used in the channel sounders, such as a PSK signal modulated by PN sequence [Prop], [KSPH01], [ZeTS04], a multitone signal with low crest factor [RUSK], and a chirp signal [SFIH02], [LPEO02].

All the signals are designed to have small envelope fluctuation to reduce the backoff of the TX high power amplifier. VNAs use step frequency Continuous Wave (CW) signal. Hybrid stepped frequency and PN sequence is used to realise very wide bandwidth [KaWe00], [VLSK03].

The phase and timing at both sides must be synchronised by some means. When TX and RX are close in distance, the local signal can be shared via coaxial cable

or optical fibre. Otherwise, high stability frequency standards such as rubidium oscillators are usually used for the purpose. For high accuracy timing synchronisation, use of Global Positioning System (GPS) and Transmission Control Protocol (TCP)/Internet Protocol (IP) connection have been also considered [LPEO02]. Frequency division multiplexing of the reference signal and the sounding signal is an alternative technique [MATB02].

To obtain the single-directional information, an array antenna is deployed at the receiver side. Time domain multiplexing using a high speed RF switch and a single RF front end is the most popular architecture. Alternatively, parallel receivers can be deployed [SFIH02]. This is advantageous for the measurement of large Doppler shifts due to the shorter measurement time. A multiport network analyser can also be considered as a parallel receiver [MGPCBRJL04]. Yet alternatively, a synthetic array approach can be used [MATB02], [VLSK03], [HaTa03b]. Here the measurement must be performed in a static environment due to the time consuming physical motion of the antenna elements. However, the SISO sounder hardware can be used without modification, and no array antenna calibration is necessary.

For the double-directional measurements, array antennas must be introduced at both TX and RX. Time domain multiplexing can be deployed at the both TX and RX [KSPH01], [Prop], [RUSK]. The switch timing must be synchronised between the TX and RX multiplexers. As the TX multiplexer handles high power, it operates slower than the RX multiplexer in general. Therefore, the switch timing of the multiplexers is designed so that the RX multiplexer scans the array for each TX antenna element. Some hybrid techniques are also possible solutions, such as time domain multiplexing at TX and parallel RXs [RaSa04], and synthetic array at TX and time domain multiplexing at RX [Herd04]. A full MIMO transceiver is also possible, by using orthogonal signals such as code division multiplexing [ZeTS04], [ChSA04] and frequency division multiplexing [SaTA02].

Antenna architectures

Antenna elements

For the non-directional measurements, omnidirectional antennas have been preferred, such as dipole antennas. The antenna element of the synthetic array may also be an omnidirectional antenna due to its original non-directional use. Slot antennas can be used as horizontally polarised omnidirectional antennas [Rich05].

Patch antennas are often used as directional antennas deployed in the base station side, or as elements of conformal arrays to avoid the influence of the support structure of the array. Patch antennas are also advantageous for polarimetric measurements, as the dual polarised element can be easily realised by using two orthogonal feeds in a single patch element [KSPH01], [Rich05], [SLRT03].

Array geometries

Array geometries are designed so that necessary angular information can be obtained. To satisfy the sampling theorem, the element spacing should normally be designed as less than half wavelength. Otherwise, there may appear grating lobes causing estimation errors of directions of arrival/departure.

Azimuth-only measurements can deploy either a uniform linear array [MATB02], [VLSK03], [SLRT03], [RaSa04], [Herd04], or a Uniform Circular Array (UCA) [SFIH02], [Rich05], [RaSa04]. With the same number of elements, the former has higher resolution in the broadside and is suitable as the BS antenna, while the resolution of the latter is insensitive to the direction and is suitable as the mobile station antenna. The linear array has another disadvantage that it suffers from cone ambiguity and the elevated direction of arrival results in an error of the azimuth angle estimate. To reduce the angular dependency of the resolution, use of a cross array antenna is also considered [VeLC03a].

For joint azimuth and elevation measurements, use of a planar or a solid structure is needed. A UCA can also be used for this purpose [HaTa03b]. Alternatively, a uniform rectangular array may be used [HaTa03b], [Herd04]. To avoid the ambiguity between angles below and above the horizontal plane, a solid array structure is necessary. A spherical array [KSPH01], [KKVV04], a circular cylindrical array [TFBN04], and a rectangular solid array [TsHT04] have been deployed.

Array calibration is necessary for the real array antenna, and the array antenna response must be measured in the anechoic chamber in advance. A continuous reconstruction technique of the array antenna response from discrete data has been proposed in [LaRT03]. On the contrary, the theoretical array response can be used for the synthetic array antenna since there is neither mutual coupling between the antenna elements nor variation of properties for individual elements. When the uniform linear or rectangular array is deployed, ESPRIT algorithm can be applicable.

There has been a trial to optimise the array geometry by using a Genetic Algorithm (GA) [KaKC02].

4.7.4.3 Parameter estimation

As explained in the previous sections, the design of adaptive arrays or smart antennas brings several technical challenges. It is clear that the knowledge of the DoA parameters is very important. High resolution techniques can resolve these parameters very accurately, which is a prerequisite for the design of antennas. In this subsection, promising methods for determining the full

double-directional and polarimetric parameters of the mobile channel are given.

The high resolution parameter estimation techniques are divided in three different types. Conventional methods are based on classical beamforming techniques and require a large number of antenna elements to achieve high resolution. The second class of methods are the subspace methods and are high resolution sub-optimal techniques, which exploit the eigenstructure of the input data matrix. The MUSIC technique for instance uses the eigenstructure of the covariance matrix. In contrast, the ESPRIT exploits the rotational invariance of the underlying signal subspace. The last class of methods is the ML technique, which also performs well under low signal-to-noise ratios. The SAGE algorithm is a way of ordering the maximum likelihood approach in an appropriate way.

This subsection is organised as follows. First, new results of the MUSIC algorithm will be shown, followed by the reporting on the use of the ESPRIT technique. The most common technique with high resolution is of course the SAGE algorithm. Theoretical and experimental results of the implementation of the algorithm in the time domain will be shown. Also the implementation of the SAGE in the frequency domain with important adaptations will be demonstrated with simulations and measurements.

The MUSIC algorithm is a high resolution signal parameter estimation technique. It exploits the eigenstructure of the input covariance matrix, based on a geometric view of the signal parameter estimation problem. Hence, it provides information about the number of incident signals and the DoA of each significant signal.

In order to decrease the computational complexity and to increase the resolution, some improvements to the previously described MUSIC algorithm can be made. The theory of the Smart MUSIC algorithm is described in [Taga97] and some practical simulations in [TaSh96]. Here, the eigenvalue decomposition is interpreted in terms of a Hermetian mapping. Using the Gram-Schmidt orthogonalisation, the speed of the data processing is ameliorated, resulting in a possibility of following the rapid change of the radio environment. This MUSIC algorithm is not only used as DoA estimation in mobile communication, but can also be used in other scenarios; [SICW01] describes the use of the MUSIC algorithm as a signal processing tool for a three-dimensional optical imaging system, based on stepped frequency radar techniques.

Another subspace-based DoA high resolution technique is the ESPRIT. It reduces the computational requirements significantly when compared with MUSIC as an extensive search through all possible steering vectors is not needed. A practical implementation of the ESPRIT is made in [BHHT04b], where the spatial scanning is configured to resemble an antenna array, also called a synthesised uniform rectangular array. With the described measurement set-up, the multipath characteristics of non-specular wave scattering from three-dimensional building surface roughness can be determined, showing that multiple paths can be detected at many scatterers. Here the delay, directly estimated by using ESPRIT, yields close agreement with the delay of the propagation path determined by using DoA estimated by ESPRIT and free space velocity.

In [SeKa01], two-dimensional fading forecast of time-variant channels is explained. Because of the implementation of a direction-of-arrival scheme as the ESPRIT technique, the channel degradation due to small-scaling fading effects must be available. Taking into account the processing delay caused by the information feedback between receiver and transmitter and the adaptation process itself, the channel variations must be predicted reliably into the future as far as possible. In this paper, the feasibility of a 2-dimensional fading forecast is studied in the context of the ESPRIT.

Next, the SAGE algorithm is investigated. The SAGE optimisation algorithm is used to replace the high-dimensional optimisation procedure necessary to compute the joint maximum likelihood estimate of the parameters by several separate maximisation processes, which can be performed sequentially.

Here, a technique derived from the ML principle is proposed. It allows for high resolution determination of the complex signals. The EM algorithm updates all parameters simultaneously, which implies a slow convergence and a difficult maximisation. The SAGE method updates the parameters sequentially by replacing the high-dimensional optimisation process necessary to compute the estimates of the parameters, by several separate, low-dimensional maximisation procedures, which are performed sequentially. These separate maximisation procedures are linked up in the final algorithm. The implementation of the algorithm can be made in the time domain and in the frequency domain. The first papers will deal with the time domain implementation.

In [VeLC04a], the simulations are done in three dimensions, resulting in the calculated parameters: the relative delay, the azimuth angle, the elevation angle and the complex amplitude of the waves. Based on these numerical simulations, convergence, resolution and performance analysis are calculated. The ambiguity of detecting different users in a three-dimensional case with a two-dimensional antenna configuration can be resolved by using for instance two parallel planes. This solution suffers from the spatially limited radiation pattern of the array elements. By using spherical arrays, there will always be some elements radiating into the direction of the signals, hence increasing the signal-to-noise ratio of the subsequent digital processing. To minimise the coupling and hence to increase the decorrelation between all the antenna elements, they should be spaced uniformly

over the sphere. This configuration will also increase the resolution of the antenna array.

Of course, using different antenna elements causes the radiation from one element to couple to its neighbours. The real current on each array element is the sum of the values due to the excitation plus all contributions from the various coupling sources from each of the neighbours. This mutual coupling is in most cases not desired, but will often be a significant factor in the total radiation characteristics. In [VeLC02a], this mutual coupling is integrated in the SAGE algorithm and the convergence and performance analysis are calculated.

Measurements are often performed in the frequency domain. To use the time domain SAGE algorithm, a transformation of the transfer function in the frequency domain to the time domain received signal, with the training sequence included, is given in [VLSK03]. The DoA parameters of the indoor measurements in a classroom are extracted with the time domain SAGE algorithm and compared with a RT tool. The same approach – evaluating measurement results in the frequency domain with a time domain algorithm – can be found in [MoRo04], where the results are also compared with a RT tool.

In fact, most measurements are broadband and the Broadband Time Domain-SAGE (BTD-SAGE) in [VeLC04b] combines the signals in another way, adapting the algorithm to be capable of detecting different paths in a broadband environment and with large antenna arrays, which is the case in the earlier described measurements. The extraction of the measurements is compared with results of a RT tool, proving that the adaptation of the algorithm works fine.

Another tool is described in [HaTa03a]: the UWB-SAGE algorithm. This algorithm divides the measured data into individual ray paths and estimates the directions of arrival, the propagation time and the variation of the amplitude and phase during the propagation for each signal. Simulations prove the quality in simple indoor propagation environments.

Also the signal model can be extended. In [VeLC03b], for the evaluation of the extracted values, the results can be compensated for the influence of spherical waves. Different simulation parameters are discussed: the extension of the measured bandwidth, the frequency resolution, the geometrical antenna array size and the number of antennas. With an appropriate choice of the measurement geometry and parameters, one can conclude that the SAGE algorithm can compensate for the properties of spherical waves. This knowledge is taken into account in [VeLC03a], where measurements with different antenna array configurations are investigated and compared to each other. Measurements of one linear antenna array that is perpendicular to the LoS and measurements of another one that is parallel to the LoS are combined to virtual measurements with a Mill's cross antenna array, resulting in better extraction results of the measurements.

In most measurements, not only the type of antenna, but also the position and the polarisation of the antenna array, are very important. By taking into account both position and polarisation in the radiation pattern of an antenna array, another adaptation to the commonly used time domain SAGE algorithm is made. Hence the direction-of-arrival parameters can be determined very accurately. Simulation results with a cubical or hexahedral antenna array are shown in [LVTC04]. Results with a dodecahedral configuration are given in [VeLC05].

Some ameliorations are suggested for the initialisation and the search procedure in the SAGE algorithm. In [StFJ02] the Initialisation and Search Improved SAGE (ISIS) algorithm is presented, which applies a bin-search-based approach in the initialisation of the parameter estimates. This modified scheme can estimate propagation paths with small amplitudes, and thus the overall profile of the environment can be thoroughly extracted. Paper [FlSJ02] reports the performance of the ISIS algorithm in experimental investigations of dispersions in delay, direction of departure and direction of incidence in a typical non-LoS environment (see Figure 4.7-28). The results demonstrate the high potential of the technique for getting a detailed insight into the mechanisms by which the electromagnetic energy propagates between the transmitter and receiver sites. In [FYSJ03], the ISIS algorithm is extended to include estimation of the polarisation matrix of individual propagation paths. This new scheme allows for estimation of the directions of departure, directions of arrival, propagation delays, Doppler frequencies, and polarisation matrices of individual paths propagating between the transmitter and the receiver sites. Experimental investigations in a non-LoS pico/microcellular environment show that the polarisation characteristics of individual propagation paths can be related directly to the types of interaction that the waves experience along their paths, such as reflection, diffraction and scattering.

An ambiguity problem in estimating Doppler frequency and directions is analysed in [YiFS03] and [PPYF04]. This problem occurs when high-resolution channel estimation methods are implemented in the channel sounding using a switched timing scheme. Theoretical analysis clarified that the switched timing schemes adopted in such systems can be constructively incorporated into the ISIS algorithm. This attribute allows performing the Doppler frequency estimation in a large scope and high resolution without the ambiguity problems. Experimental investigation using PROPSound illustrated this merit in mobile radio environments. Report [HaTa03b] presents a MIMO sounding system dedicated to short range wireless communication. Using the system, a double directional measurement campaign

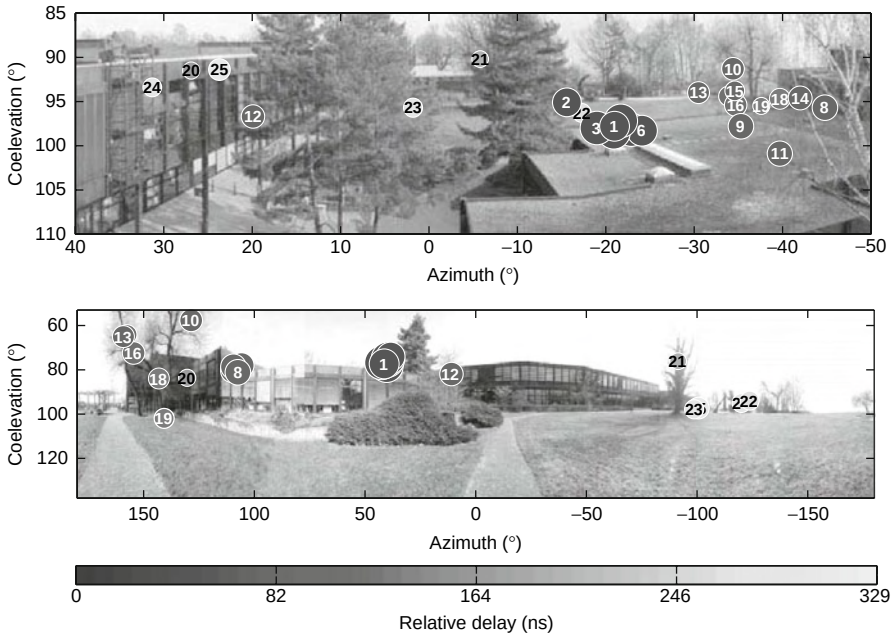


Figure 4.7-28 Experimental data obtained with the ISIS algorithm.

in an indoor NLoS environment was performed, where the channel was estimated in a deterministic way with the above-mentioned ISIS algorithm. Estimation of polarisation characteristics for identical paths was proposed.

An approach leading to lower complexity and faster convergence of the SAGE algorithm is derived in [RiOK04]. The proposed method employs the Von Mises distribution model defined in the angular domain. This models the directional data observed in channel measurement campaigns using only few parameters. The mixture model allows for representation of scenarios where multiple scatterer clusters are present, and also for representation of each cluster as a finite mixture of Von Mises distributions. The benefits of this method are due to smoother likelihood function and significantly lower dimensionality of the model, leading to lower variance of estimates and reduced computational complexity.

Another approach with the SAGE algorithm is the implementation in the frequency domain, which is more straightforward with frequency domain measurements. An overview of various maximum-likelihood algorithms with different implementation methodologies to estimate multipath parameters can be found in [TaBN04]. Due to the iterative nature of the algorithms, and with the requirement of estimating multidimensional multipath parameters from a vast measurement database, a number of simplified solutions has been proposed in this paper. The performance of the proposed algorithms is compared with that of the classical SAGE algorithm. One of the described algorithms is the Hybrid-Space SAGE (HS-SAGE) algorithm, which is a combination of the element-space and beamspace processing. In [PaTB04], a capacity-based analysis of directional MIMO channel measurements is presented. It is not only suitable for use with a circular array, but it also enhances the processing speed without sacrificing the accuracy and the resolution. A comparison of directly measured MIMO channels and MIMO channels generated from the multipath parameters, extracted using the HS-SAGE algorithm, is described to prove the previously mentioned statements.

In [ThLR04], an ameliorated version of this frequency domain SAGE algorithm is implemented. Instead of the normal serial interference cancellation, the parallel interference cancellation is suggested. The theory is checked with measurement results.

Another implementation of a high-resolution spectral estimation scheme, based on the SAGE algorithm, is suggested in [SeKa02a], and is compared with the subspace-based ESPRIT algorithm. By applying high-resolution parameter estimation schemes on time-invariant channels, the actual and past channel information is analysed to predict the future channel situation. For time-variant situations, parameter tracking or parameter estimation takes the evolutional nature of the radio channel into account. Applied on RT simulation data, a fading forecast using these different estimation schemes is calculated. A toolbox for spectral analysis and linear prediction of stationary and non-stationary signals is described in [Semm04a]. It includes the generation of test signals, the spectral analysis and parameter estimation of the test signals and measurements and the linear prediction of the signals.

Polarisation-sensitive antenna arrays at both link ends (mobile station and BS) of a MIMO channel are included in the parameter estimation with the SAGE algorithm of

[Rich05]. The joint estimation of the four complex polarisation transmission path weights (the directions of arrival, directions of departure, time delays and Doppler frequencies) of measurements showed the polarisation dependence of wave propagation in a mobile radio channel.

An extended data model for high resolution channel parameter estimation and parametric channel modelling is introduced in [TLRT05b]. This is totally different from the well-known ray-optical-based data models, which contain only discrete specular propagation paths. In this paper, distributed diffuse scattering components are also included. For these parameters an estimator is developed and their Cramer-Rao Lower Bound (CRLB) is derived. Finally, the integration of this extended channel model into the SAGE algorithm is discussed.

An newly introduced multidimensional maximum likelihood parameter estimator is described in [ThLR04] and is called the RIMAX algorithm. It estimates jointly the parameters of the specular components (propagation paths) as well as the parameters of the distributed diffuse scatterers. The algorithm is based on the conjugate gradient optimisation strategy. In comparison with the SAGE algorithm, the complexity is drastically reduced, especially if the number of paths (components) is small compared to the number of observations. Additionally, the algorithm provides an estimate of the variance of the calculated parameters, yielding reliability information of the channel parameters estimated. This tracking algorithm has been tested in a static SIMO scenario in an urban environment and described in [AlRT04].

In a MIMO indoor propagation environment, the knowledge of cluster characteristics is very important. Here also the SAGE algorithm proves its qualities. First, identification of clusters was done visually using the double-directional angular power spectrum. The characteristics of each cluster were calculated with the new algorithm, based on SAGE estimation of eigenmodes. The next step in the development of RIMAX is proposed in [RiEK05]. A state-space-based method, describing the dynamics of the channel parameters in the time, for parameter tracking in order to estimate the propagation paths is suggested here. This results in a lower computational complexity and allows the tracking of the desired time varying parameters as well. The performance of the proposed technique is investigated using channel sounding measurements.

4.7.4.4 Device performance and reliability of channel structure estimates

A thorough investigation of the multidimensional wave propagation mechanisms is a prerequisite for understanding the spatial and temporal structure of the channel transfer matrix. The accuracy and reliability of the estimated channel parameters from a field measurement are dependent on the channel sounding system design. In case of the spatial domain the design of the antenna arrays [TLAP02], [KaKC02] is one parameter which determines the reliability of the DoA estimates. Due to practical aspects, e.g. mechanical, electrical construction and imperfections of the antenna arrays, the performance of real antenna array differs from the designed and simulated one. Therefore methods for calibration and evaluation were developed [LaRT04a], [LaRT04b].

The proposed design and the evaluation methods of antenna arrays are based on techniques using the CRLB and beamformer optimisation. The CRLB defines variance limits of the DoA estimates independent from any parameter estimator. It results from the calculation of the inverse covariance matrix of the matrix of the first-order derivatives of the observed data with respect to the channel parameters (e.g. DoA). Therefore a derivable data model of the antenna array is required. For this purpose analytic functions of the array manifold are often used in case of design and simulations [KaKC02]. These approaches often neglect the influence of the antenna beam patterns, mutual coupling and other characteristics or imperfections of real antenna arrays for simplicity of the derivatives. Therefore a derivable data model based on calibration measurement data of real antenna arrays is proposed [LaRT04a].

The methods using the CRLB to design an antenna array assume that the estimation algorithm will find the global minimum of the cost function of the optimisation problem. Most of the estimation algorithms are sequential; this means that they estimate the channel parameters path by path and parameter by parameter. This can result in a solution which is just a local minimum of the cost function (e.g. estimation of phantom paths). Therefore also an optimisation of the beamformer is considered [TLAP02]. Both the CRLB minimisation step and the beamformer optimisation step should be combined to achieve satisfying results using the designed antennas in channel measurements and analysis.

CRLB minimisation step

One approach for the optimisation of antenna arrays used in mobile radio channel characterisation devices is to apply a GA [KaKC02]. It is proposed to design the array of a channel characterisation system, in order to minimise the CRLB (minimum variance) of the DoA estimates. A critical issue is to choose the optimal array geometry, namely to find the positions where the array elements should be placed, in order to obtain measurements that reduce the DoA estimation error to a minimum, assuming a minimum variance unbiased estimation algorithm, e.g. ML algorithm. Each sensor is supposed to have an isotropic beam pattern. It is shown that the GA determines

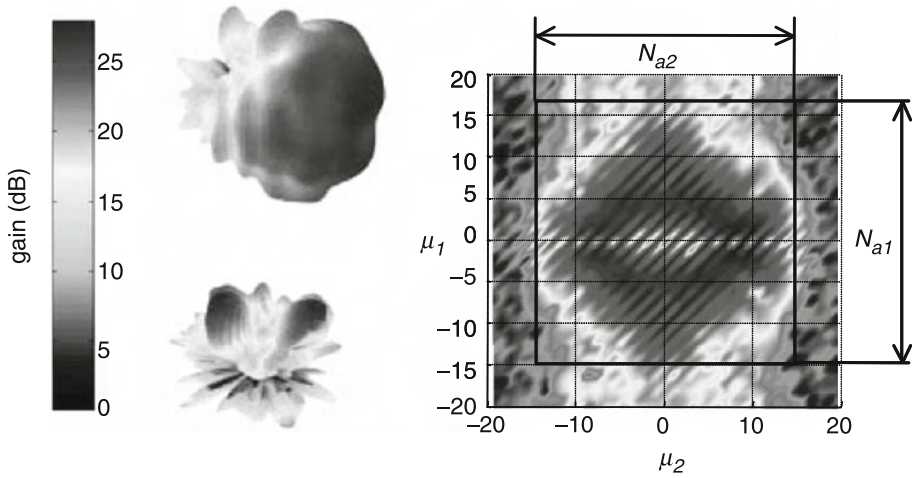


Figure 4.7-29 Polarimetric beam pattern for vertical (V) and horizontal (H) excitation (left) and corresponding EADF (right) for the vertical excitation (VV) of an SPUCPA.

the position of the sensors to minimise the mean CRLB of the DoAs for unknown number and position of scatterers and also in case of prior knowledge of the clusters from which the signals originate.

Beamformer optimisation step

To avoid estimation results which are local minima of the cost function in [TLAP02] it is proposed to suppress the side lobes of the beamformer $c_k\,(x, \Theta) = x_k^H a\,(\Theta)$, where x_k is the complete data of the kth path and $a(\Theta)$ is the normalised steering vector of the antenna array with respect to the DoA Θ. In the initialisation step, from most ML algorithms such as EM and SAGE this function is calculated. For an example of a UCA with omnidirectional beam patterns the best side lobe level suppression is achieved by a radius of: $r_{opt} \approx N\lambda/16 < \lambda/4\sin(\pi^2/N)$. Further suppression is possible through the joint optimisation of the beam pattern of the antenna elements and the array radius. It is found that the radius of the UCA should be as big as possible in terms of a minimum CRLB of the DoA. In terms of side lobe suppression of the beamformer a smaller radius is preferred. Therefore it is necessary to find a compromise between the minimum CRLB and the optimum beamformer.

Evaluation of real antenna arrays

Once the antenna array is realised it can be measured in an anechoic chamber. The antenna response is described by the complex polarimetric beam patterns of all elements which result from vertical and horizontal polarised excitation. Recording of the complete spherical beam pattern requires precise rotation of the array around a suitable defined pivot point located in the phase reference centre of the array and excellent phase stability of the set-up. The measured beam patterns are discrete in azimuth and elevation. Due to the periodicity of the beam patterns in 2π the DFT transforms beam pattern to the Effective Aperture Distribution Function (EADF) domain [LaRT04a], [LaRT04b]. The EADF matrix is calculated by a 2D Fourier transform by applying the Fourier matrices of the azimuth and the elevation to the discrete beam patterns. The EADF concept allows a considerable data compression since it is distinguished by a limited support area as shown in Figure 4.7-29 (right). On the other hand, it allows a simple analytic calculation of the derivatives with respect to the angular parameters.

Once one has the first-order derivatives with respect to the DoA the Fisher Information Matrix (FIM) and its inverse the CRLB of an arbitrary test scenario can be calculated. The investigation and exploitation of the FIM structure is essential to design a robust and efficient parameter estimator. The simplest scenario is a single impinging path scenario, which is analysed in the following example. It is related to an SPUCPA with 192 elements. In this scenario the FIM is diagonal. In Figure 4.7-30 the CRLB of the azimuth angle is compared to the variance which is achieved by an experiment. For any true azimuth/elevation pair φ, ϑ within the coverage sector of the SPUCPA the azimuth was estimated by an ML procedure [Rich05], [ThLR04]. This experiment was repeated 64 times. The noise level for CRLB calculation was adjusted to match the observed device noise level which was held constant according to a SNR in the main beam direction ($\vartheta = 90$ deg) of 17–18 dB. It is also shown that this evaluation method is valid in the case of more complicated scenarios.

Further, the described method can be used for the ML parameter estimation as clarified in [TLRT05b], [ThLR04], [Rich05]. Here it is used to define the fundamental limit of the achievable variance of the estimated DoA/DoD parameters in terms of the CRLB to

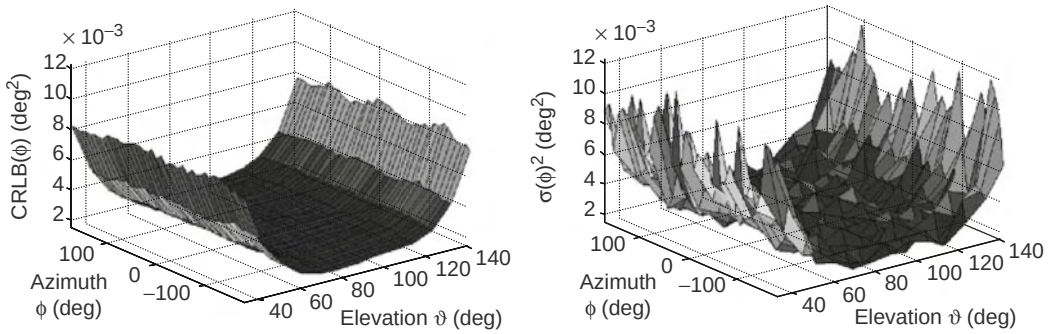


Figure 4.7-30 CRLB (left) and estimated variance (right) of estimated azimuth DoA vs. the true azimuth and elevation direction range.

check the reliability of the obtained parameters during the estimation process.

Another approach for evaluating the performance multidimensional channel sounding is presented in [MuSV05]. There the idea is to use existing already parametrised measurement data to create a dataset that resembles closely the original channel information. The approach is based on combining the radiation patterns obtained in calibration measurement of the measurement arrays with the parametrised propagation data. This way a well-known but still realistic test data with multiple signals can be created for studying, e.g. the correctness of estimated angular distributions.

4.7.5 Channel characterisation

4.7.5.1 Introduction

This section will mainly deal with aspects of the variability of power received through the mobile radio channel. These aspects comprise the dependence with distance, both on a small and large scale, and its temporal and angular dispersion, all regarded separately.

The structure will be as follows:

1. a short introduction to the formal requirements on the processing of measured data;
2. a subsection on path loss, its distance dependence and small- and large-scale fluctuations;
3. a subsection on temporal and angular dispersion; and
4. a subsection on fading prediction techniques.

4.7.5.2 Statistical processing of measurements

The properties of the mobile channel to be treated in this subsection are all modelled as random processes. The bulk of the characteristics are derived, as in the majority of the literature on channel characterisation, from operations on time series. In this way, the following is implicitly assumed [Bult04]:

- the measured variability of the random process under consideration is only a function of either time or distance, making it a stochastic process;
- the variability is only a function of the time difference or distance interval between two samplings, as in a Wide-Sense-Stationary (WSS) stochastic process. As a result, (ensemble) mean values should be independent of time or distance and autocorrelation functions should only depend on the time or distance difference;
- the stochastic process is ergodic, allowing replacing ensembles by isolated sample functions, mostly time series.

Effectively, this means that the mean of the stochastic process should be constant and that the covariance of the process should exist, conditions that could prove difficult to test in practical situations. It is also fair to say that in the literature tests are seldom reported that justify the above assumptions.

One possibility is to assure that the stochastic process is both WSS and Gaussian, as such processes are ergodic. However, Gaussian properties should be inferred through the central limit theorem, as time series analysis inherently assumes ergodicity. Bultitude takes the occurrence of less than 15 multipath components on the radio channel as an indication that Gaussian properties cannot be assumed for the channel transfer [Bult02b]. Depending on the distributions and the independence of the multi-path components, much larger numbers than 15 could be needed to safely assume Gaussian properties.

Bultitude tested on WSS [Bult03] by comparing subsequent Doppler spectra. He used a MUSIC algorithm to extract Doppler parameters from half-overlapping 1 metre segments of measurements and computed the running variance of both the mean power-weighted AoA and the angular spread. Changes of greater than 10% of the maximum value of either variances were taken to indicate non-stationarity. For measurements in a microcellular

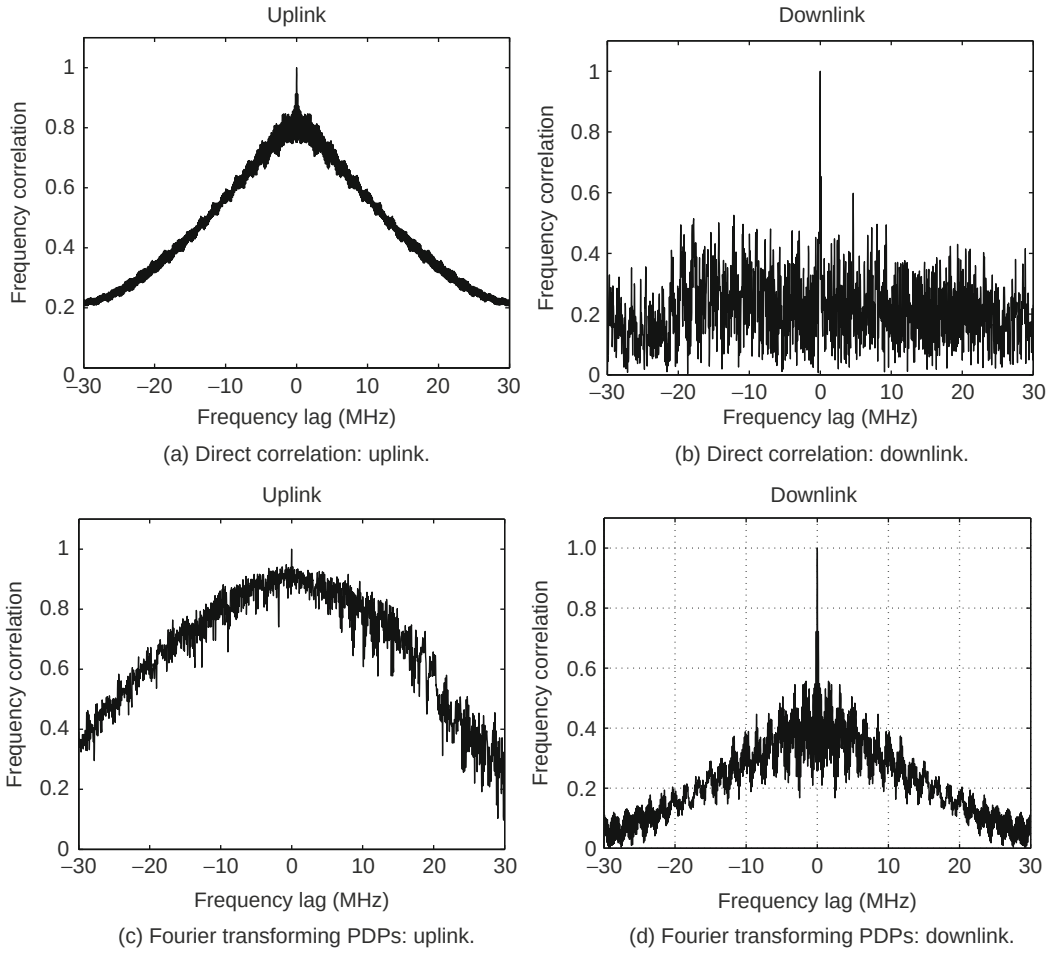


Figure 4.7-31 Comparison between frequency correlation estimates at 2 GHz (after [SaGo01]).

scenario in downtown Ottawa (Manhattan grid) at 1.8 GHz, approximately negatively exponentially distributed stationarity intervals, 'consistency lengths', were found, but the measurements showed greater probabilities for longer intervals. For NLoS streets, the median values of the lengths of these stationarity intervals ranged from 2.6 m to 3.3 m, depending on street orientation [Bult03]. Another technique developed by Bultitude for testing on WSS is inspection of the variation around the mean in spectral lines [Bult02b].

In the process of determining frequency correlation functions, it is common practice to use a Fourier transform on the APDP. However, this assumes Uncorrelated Scattering (US). One test on US is to determine frequency correlation functions directly from the spectrum and to test whether the resulting correlation functions are symmetric with respect to arbitrary reference frequencies [Bult02b]. The validity of the assumption of US in indoor environments has been treated by Kattenbach [Katt97]. Also, the presence of dominant coherent components, as in Ricean fading processes, can distort results obtained by Fourier transforms. Salous and Gokalp presented an example where the relatively small differences between urban up- and downlink characteristics at 2 GHz resulted in clear differences between frequency correlation estimates despite similar power delay profiles [SaGo01]. Figure 4.7-31 shows these differences between up- and downlink for frequency correlation estimates by direct correlation as opposed to those derived from Fourier transforming power delay profiles.

4.7.5.3 Path loss and building penetration

A vast number of different path-loss models have been proposed in the past two decades, ranging from simple empirical to more complex physical ones [Corr01]. Each of these models has it owns advantages and drawbacks. Hence, path loss in outdoor macro- and microcells is addressed, including vegetated residential environments. Indoor and building penetration models are also presented.

An approach to extend the validity domain of simple empirical models, e.g. Okumura-Hata [OOKF68], [Hata80], in urban areas is proposed in [GCBV02]. It is assumed that propagation in urban environments occurs by means of reflections and diffractions either along the streets (lateral propagation) or over the buildings (vertical propagation) therefore, two independent models can be combined into an overall statistical one. The two components are treated separately, and each of the submodels can be refined independently of each other. The obtained model leads to the well-known exponential law when BS antenna height and path length are large enough, the Okumura-Hata model is being used in its validity domain, and is being replaced by the proposed extension for distances below 1 km and for BS antennas lower than 30 m. The mix between vertical and lateral propagation, as well as the existence of LoS, is expected to change as a function of the distance. A dominant mechanism exists at any distance from the BS, which allows the identification of three regions, each one characterised by one of these mechanisms.

Once these regions are identified, it is necessary to identify the transition modes between them, and the weighting factors to use for each of the three components. In order to identify the region where LoS propagation is dominant, an obstruction probability factor, p, is used, defining the probability that a link between the BS and the MT is obstructed. Globally, p will be zero in the vicinity of the BS, and it will increase with distance, approaching unity. The path loss in an urban environment is obtained as

$$L_p = (1-p)L_p^{LoS} + pL_p^{NLoS} \tag{4.7.4}$$

with L_p^{LoS} and L_p^{NLoS} being the path loss in LoS and NLoS, respectively.

The NLoS component results from the combination of two terms, representing the vertical and lateral propagation. For simplicity, it is assumed that the total field strength is given by the strongest component, thus,

$$L_p^{NLoS} = \min(L_p^{LAT}, L_p^{VERT}) \tag{4.7.5}$$

where L_p^{LAT} and L_p^{VERT} represent the lateral and vertical attenuations, respectively.

In order to apply the model, three components need to be evaluated as a function of the geometrical parameters of the propagation scenario.

Assuming that the received field strengths in the lateral and vertical planes are statistically independent, the total received field strength standard deviation can be evaluated from (4.7.6)

$$\sigma(d) = \sqrt{\sigma_{VERT}^2(d)q(d) + \sigma_{LAT}^2(d)[1-q(d)] + [\mu_{VERT}(d) - \mu_{LAT}(d)]^2 q(d)[1-q(d)]} \tag{4.7.6}$$

where q is the dominance probability, i.e. the probability that, at a given distance, d, the dominant component in the received field is the one due to vertical propagation, and the mean value and standard deviation for the two different components are obtained as $\mu_{LAT} = \alpha_1 d + k_1$, $\mu_{VERT} = 10\alpha_2 \log(d) + k_2$, $\sigma_{LAT} = \min(\beta_1 d, 20)$ and $\sigma_{VERT} = \beta_2$.

Parameters α_1, α_2, β_1, β_2, k_1 and k_2 depend on the geometrical and topological characteristics of the considered area, e.g. mean building height, mean street width, BS antenna height, etc. Values of $\alpha_1 = 0.108$, $\alpha_2 = 7.324$, $\beta_1 = 0.027$, $\beta_2 = 6.783$, $k_1 = 64.777$ and $k_2 = -70.000$ were obtained from fitting of measurement data from a small cell at 947 MHz. Additional data for 1800 MHz can be found in [BCFF02].

It is known that path loss depends on the carrier frequency. Wideband measurements at the same environments in two frequency bands, namely 2.45 and 5.25 GHz, and a system bandwidth of 100 MHz, are presented in [LZMJ05]. Rural measurements were performed in a small rural municipality, Tyrnävä, which is located close to Oulu, Finland. The measurement route was about 1 km long and mainly LoS type, thus being representative of flat rural LoS macrocellular environments. The suburban environment is a typical suburban residential area in Oulu region with at most four to six-storey buildings in the vicinity (one- to two-storey buildings in the measurement route), and not so wide streets. The street grid is rather regular, and no open areas can be found in between the houses. An unobstructed LoS condition was observed. Macro- and microcellular cases were considered by positioning the receiver at different heights, 11.7 and 7.6 m being considered, respectively.

Based on measurements, a simple log-distance path-loss model is derived by using linear regression of the scatter plot of the measured data. The results for rural and suburban cases are reported in Table 4.7-5. Some figures for the rural case are shown as an example, Figure 4.7-32.

In [ZVRK04], empirical path-loss models for microcells are derived from measurements at 5.3 GHz and a system bandwidth of 100 MHz, performed in an urban microcellular environment in the centre of Helsinki. The environment is characterised by a regular street grid, where the building height varies between four and seven storeys. The BS antenna was placed in the middle of street crossings at height of approximately 10 m above ground level, i.e. clearly below the rooftop level. At the BS, a horizontally aligned uniform linear array of four dual-polarised microstrip patch elements was used. At the MT, two omnidirectional (dipole)

Table 4.7-5 Path-loss in rural and suburban environments.

	Lp [dB]	
f **[GHz]**	**Rural**	**Suburban**
2.45	$38.3 + 21.1 \log(d)$, $d \leq 650$ m $-105 + 75.0 \log(d)$, $d > 650$ m	Not available
5.25	$41.8 + 22.0 \log(d)$	'Macrocell' $43.3 + 22.8 \log(d)$ 'Microcell' $41.6 + 23.8 \log(d)$

elements and a mock-up user terminal with four antennas were used.

Three BS sites were chosen at different street crossings, and in each BS location several BS array orientations were measured. In each BS site, the array was always directed towards a street canyon to measure the received powers along the canyon and the perpendicular streets. Measured streets have different widths within approximately 20 to 30 m.

Empirical path-loss models were derived from measurement data by using the least square error method. For the LoS case, seven streets were measured, a path-loss exponent of 2.34 and a standard deviation of 2.6 dB being obtained. For the NLoS one, nine perpendicular streets were measured. It was observed that the path-loss exponents are within 1.6 to 2.9, with a mean value of 2.34, the standard deviation is within 1.9 to 5.6 dB, with a mean value of 3.1 dB.

Field prediction in vegetated residential environments is not easily accomplished by standard path-loss prediction models [Corr01]. The electromagnetic fields of radio waves propagating through a tree canopy can be decomposed into mean (coherent) and diffuse (incoherent) components. At low frequencies, the diffuse component is relatively small and only mean fields need to be considered

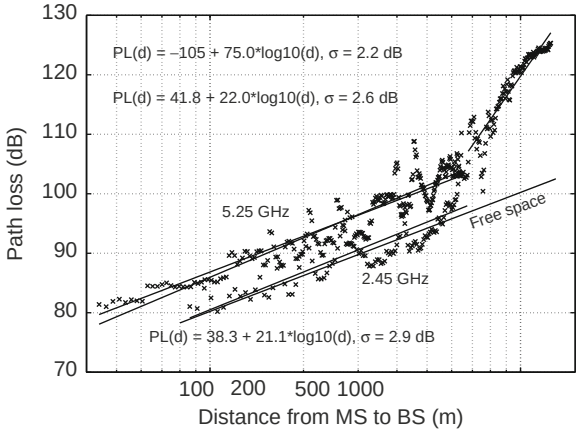


Figure 4.7-32 Rural path-loss models at 2.45 and 5.25 GHz [LZMJ05].

[ToBL98]. As the frequency increases, spatial fluctuations of the field must be taken into account. Also, even at low frequencies, the incoherent fields become important as the distance between the transmitter and the receiver increases in a random scattering medium. In [ToLa04], the behaviour of the incoherent component at Very High Frequency (VHF) and Ultra High Frequency (UHF) is examined for a trunk-dominated environment. It is assumed that trunks are symmetric, and that transmitter and receiver are in the same transverse plane. The phase function appearing in the radiative transport theory is obtained from the differential scattering cross-section of a dielectric cylinder, being valid for both low and high frequencies. The exact radiative transport equation is solved numerically by the eigenvalue technique. The solution is used to compute both the coherent and the incoherent attenuation constants in the trunk-dominated environment.

For a complex cross-section permittivity of $10^{-3}j$, a tree trunk radius of 10 cm, and a density of 1000 trees per hectare [ToLa04], it is observed that attenuation expressed in dB decreases linearly with increasing distance, the incoherent component contributing over the coherent one. For example, at 100 m the incoherent intensity adds 10 dB over the coherent attenuation. Globally, the coherent attenuation, in dB, as a function of distance, is given by $0.21d$, whereas the total attenuation is obtained as $0.11d$.

Indoor measurements at 2.45 and 5.25 GHz, and a system bandwidth of 100 MHz, carried out at Elektrobit premises in Oulu during night time in an unpopulated office are presented in [LZMJ05]. By using linear regression of the measured data the path loss was evaluated, $L_p = 48.1 + 17.3 \log(d)$ and $L_p = 3.3 + 52.1 \log(d)$ being obtained at 5.25 GHz, for the LoS and NLoS cases, respectively. At 2.45 GHz, $L_p = 42.7 + 16.4 \log(d)$ and $L_p = 10.8 + 45.4 \log(d)$, was evaluated.

A method for modelling path loss inside office buildings is presented in [MeBe02]. The proposed formulation is a combination of a street microcell model with a power law one that is valid in and around straight corridor sections. The model has been validated by comparing it to path-loss measurements in a large office building with a complex topology. Narrowband measurements were performed at 5.2 GHz. The topology of the building was quite complex, consisting of several straight corridor sections, which were connected at crossings and turnings. The building was 70 m wide, 100 m long and five storeys high. The floors were made of reinforced concrete, while the outer walls were made of brick and reinforced concrete, and the inner walls between each room were made of double plaster board supported by vertical metal studs.

The basic power law model does not account for indoor topologies with corridor segments which are connected at crossings and turnings. The recursive model, however, was designed for such topologies [Berg95]. By combining the

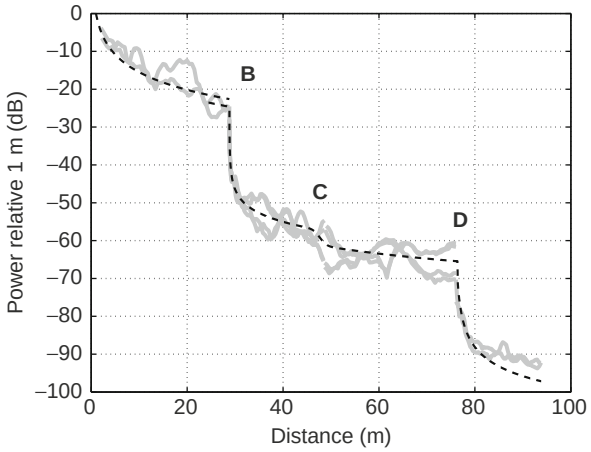


Figure 4.7-33 Received power for the corridor–corridor class [MeBe02].

power law model, and the recursive model, complex indoor environments can be easily modelled [MeBe02].

By fitting the model parameters to the measurement data, path-loss exponents of 1.55, 2.7 and 3.9 were obtained for the *corridor–corridor*, *room–corridor* and *room–room* classes, respectively.

In Figure 4.7-33, the received power is shown (solid line) relative to the power measured at 1 m from the BS, and the results of the combined model for the case of *corridor-corridor* class (dashed line). Points B and D are nodes of the recursive model, and at C a loss of 4 dB due to an obstruction is considered.

The millimetre frequency band at 60 GHz is one promising candidate to provide transmission data rates up to several hundred megabits per second, and it has been proposed for future WLANs [Corr01]. From wideband measurements in rooms and corridors of two different buildings of the Helsinki University of Technology, several results were derived [Geng05], Figure 4.7-34.

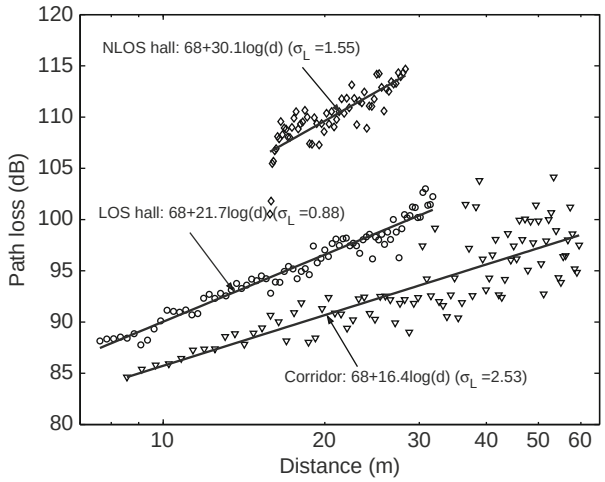


Figure 4.7-34 Path loss at 60 GHz [Geng05].

Table 4.7-6 Attenuation in hospital I at 2.45 GHz.

Room	Attenuation [dB]	
MRTo	Window	95.5
	Door	73.2
	Wall	91.5
X-ray	Door	41.5
	Wall	51.3
Operating	Wall	61.6
Plasterboard	Wall	6.4
Concrete	Wall	35.5

In contrast to usual buildings, hospitals are partly constructed of special walls with metallic layers inside. The wall structure depends on the kind of hospital room. Narrowband measurements in relevant areas of four different hospitals are presented in [ScFW02].

The results of the measurements reveal that it is important to consider the room as a whole, when determining the attenuation to neighbouring rooms. Slots and openings in walls give rise to coupling effects through rooms with metallic layers inside the wall. In order to accurately investigate wave propagation properties, diverse antennas and measurement systems were used. The results show a varying behaviour of the wave propagation in different rooms of the hospital.

The attenuation of single walls was measured by positioning horn antennas on both sides of the wall, and then measuring the attenuation of the transmitted power directly through the wall. Measurements were performed at 2.45 and 5.2 GHz, with the antennas positioned at 50 cm from the wall. In Table 4.7-6, the results of the measured attenuation of the walls in hospital I at 2.45 GHz are shown.

Table 4.7-7 shows results of wall attenuations measured in hospital II. Hospital I was newly built in 2000, whereas hospital II was built in 1970. This means that walls are built in different ways, especially those of the rooms for X-ray examinations, which are made of plaster blocks in hospital II. This leads to different results compared to that of Table 4.7-6.

A common result is that the attenuation increases for higher frequencies, with a difference ranging from about 3 dB for concrete to about 15 dB for Pb-shielded walls. Another interesting result is that the wall attenuation is generally lower for horizontal polarisation. A possible reason for this behaviour could be the vertical structure of the walls: all the slots in the metallic shielding and all the steel beams of the basic wall structure are vertically aligned.

Table 4.7-7 Attenuation in hospital II at 2.45 and 5.25 GHz, vertical and horizontal polarisation.

	2.45 GHz		5.25 GHz	
Room	Vert. pol.	Hor. pol.	Vert. pol.	Hor. pol.
Concrete (10 cm)	7.4	6.8	10.7	10.3
Operating	38.2	39.9	52.7	48.3
X-ray wall	42.6	31.1	61.9	46.6
X-ray door	44.2	30.4	61.7	56.9
MRTo wall	91.4	81.2	95.7	95.3
MRTo door	73.7	67.4	93.7	102.9

There are mainly two different approaches for indoor coverage planning:

1. picocells deployment, where the BS is installed inside the building, e.g. Digital Enhanced Cordless Telecommunications (DECT), and
2. indoor coverage from an externally installed BS (micro- or small macrocells deployment), where, in most cases, propagation takes place over NLoS conditions.

A great deal of the radio coverage inside buildings is still being provided with the use of BSs located outdoors. The building penetration loss is defined as the difference between the average signal strength in the local area around a building and the average signal strength on a given building floor. For radio planning purposes, the building or room loss factor can be used as an addition to the predicted signal loss for the surrounding local area. At higher floors, the received signal strength will be higher than at the ground floor. This is important for radio planning, because it may cause higher interference levels at higher floors.

Empirical penetration loss models for indoor coverage planning and their parameters for the 1.8 GHz frequency band, experimentally supported on a measurement campaign that has been performed inside 11 different buildings in the city of Lisbon, are presented in [XaVC03]. The models' reference is the propagation loss evaluated at an external wall 1.5 m above the ground level.

The measured buildings were divided into four classes, according to their height and type of integration in the urban structure: High Isolated, Low Isolated, High Integrated and Low Integrated. The difference between integrated and isolated buildings is made by the fact of sharing (or not) walls with the surrounding buildings. An integrated building always shares walls with some others, while an isolated one does not share any. Concerning the height parameters (High/Low), buildings are classified by the number of floors, up to four floors (included) being considered Low, else being classified as High. A description of the measured buildings' characteristics can be found in [XaVC03].

By processing the collected data, it was noticed that all attenuation histograms had a common shape for each building, thus motivating the creation of a new model, the 'Double Sided Gaussian Model'. Histograms had the shape of a kind of 'unbalanced' Gaussian function, with distinct standard deviations for the mean value of 'right' and 'left' sides. The Double Sided Gaussian Model output parameters are the mean value and the standard deviation for both semi-Gaussians, Table 4.7-8.

For the same buildings, the comparison between 1.8 GHz measurements and 900 MHz measurements was made, a difference of about 6 dB being found, hence verifying the attenuation dependence of 20 log(f). Following the same law, it is expected that the presented results can be extrapolated for the UMTS frequency bands.

In [MaHe03], measurement results taken in office buildings situated in The Hague are presented. Measurements show that long-term fluctuations can be observed between signal levels received in different parts of a building. While on lower floors these fluctuations follow a log-normal distribution, significant differences have been observed at higher floors between LoS and NLoS areas. Therefore, the relationship between the floor height and extra gain with respect to the ground floor level is not linear, depending on factors such as the radiation pattern of the BS antenna and the local urban clutter effect.

Additional results on indoor coverage by outdoor BSs at 1.8 GHz can be found in [KüMe02], empirical factors for building penetration and height gain from extensive measurement campaigns for the 1.8 GHz frequency band being provided.

4.7.5.4 Long-term fading

In a complex propagation environment, as usually found in mobile communication systems, the received signal results from all possible effects associated with the interactions between the propagating wave and the objects within the propagation space. From the analysis of narrowband

Table 4.7-8 Double-sided Gaussian parameters per building class.

Building class	Mean [dB]	Std. dev. right [dB]	Std. dev. left [dB]
High isolated	2.59	16.59	9.74
Low isolated	15.79	10.97	12.95
High integrated	13.27	13.92	12.37
Low integrated	12.83	9.81	6.28
All classes	10.19	13.85	13.84

measurements in different environments, it is commonly accepted that, in general, the average received signal magnitude is random and log-normal distributed around the mean distance-dependent path-loss value [Pars92].

These variations around the distance-dependent value are usually referred to as slow- or long-term fading, since the magnitude of the received signal remains approximately constant for short periods of time or space; therefore, variations can only be observed on a long-term basis, tenths of wavelengths being usually considered [Pars92]. This type of long-term variation is commonly justified by the changes in the visibility or obstruction of multipath components due to MT or scattering environment changes (movement). This phenomenon is related to the denotation of shadow fading, which is sometimes used when referring to long-term fading.

This distribution is a well accepted model for modelling narrowband fading, being frequently used for system simulation and analytical calculation purposes. Nevertheless, a different long-term fading behaviour can be found in specific environments, e.g. tunnels [Corr01].

Wideband channel modelling is usually based on the channel impulse response of the propagation channel. Contrary to the narrowband case, in which the channel response is described by one single component, in the wideband one, it is described by different components (paths) arriving at different delays, being related to distinct propagation paths. Thus, from the long-term point of view, it is common to model each path as a narrowband channel, i.e. assuming a log-normal distribution for each path [Corr01].

From wideband measurements in urban macrocellular environments [MiVV04], quite large differences in the parameters of the log-normal distribution associated to different clusters were found, depending on the route being considered; therefore, questioning if the assumption of the log-normal distribution is valid when an arbitrary environment, is considered, e.g. all measured routes. In this case, different clusters will contribute to the major part of the total received power, depending on the measured route.

In fact, the assumption that the average received signal magnitude in dB is Gaussian distributed is widespread but not straightforward [Hans02]. From measurements carried out within rooms and corridors of a building with rather solid inner walls made of brick, it is observed that when both are considered, the results will not yield a single Gaussian distribution but rather a superposition of two, i.e. a bimodal Gaussian, which is best described by their two independent means and variances. Therefore, the performed measurements indicate that the assumption of a Gaussian distributed signal magnitude is not valid, when the measured statistics are dominated by a particular scenario, which has for instance much higher amplitudes than all the other parts of the investigated environment. Globally, it is observed that the log-normality of the path loss is not valid for arbitrary scenarios; therefore, it is recommended to separately model each arbitrary scenario by decomposing it into several subscenarios for which the parameters of the potential multimodal distributions should be calculated separately.

In most propagation models, long-term fading is generated independently for different MTs; however, in network simulations it is common to have several MTs operating within a given area. Since long-term fading effects depend on the MT position, which in network level simulations are usually close to each other, a better approximation to reality should be obtained if shadowing effects are not modelled independently for each MT.

Taking into account that long-term fading effects are modelled by the addition of a Gaussian distributed random variable, this does not completely model the shadowing process. An additional aspect should be considered: shadowing is a slowly variant characteristic of the radio channel. This variation rate indicates the existence in the time domain of a non-zero shadowing autocorrelation. As mobility is included, time correlation is intimately related to the space one. In fact physical explanation of shadowing is primarily associated to position [Lee85]. Spatial correlation can be modelled as [ETSI97]

$$R(\Delta r) = 2^{-\frac{\Delta r}{d_{corr}}} \tag{4.7.7}$$

where Δr is the space shift and d_{corr} the decorrelation distance, for which typical values can be found in [ETSI97].

Computer generation of shadowing data, taking into account (4.7), can be realised in two different ways:

1. The first approach initially generates a set of n independently distributed random Gaussian variables, each corresponding to one of a set of n equally separated positions. Obviously, this results in a sequence that is uncorrelated. Next, the sequence is filtered so that the desired autocorrelation function is obtained.
2. An alternative procedure is described in [Vite94]. This procedure generates shadowing samples one by one, it is not necessary that they are equally separated. Once a shadowing sample, G_i, is produced, the next one is also a Gaussian random variable whose mean is $R(\Delta r)G_i$, and whose variance is $(1 - R(\Delta r)^2)\sigma^2$, σ being the standard deviation of shadowing values.

One major drawback of the second approach is that the autocorrelation depends only on the distance shift. Considering the case in which a series of propagation maps that account for shadowing needs to be generated, it involves producing a shadowing sample for each location in every map. Since maps are two dimensional, it is not possible to establish an order among its locations; consequently, the second approach cannot be adopted. In this situation, instead, the first one should be used; nevertheless, its application is not straightforward.

A two-dimensional modelling technique, which allows MTs in the vicinity of the simulated environment to undergo related shadowing effects, is presented in [FrLC03], [FGLM04]

$$R(\Delta x, \Delta y) = 2^{-\frac{\sqrt{\Delta x^2 + \Delta y^2}}{d_{corr}}} \quad (4.7.8)$$

where Δx and Δy are the MT shift in horizontal and vertical coordinates, respectively, and d_{corr} is the decorrelation distance.

Assuming that a set of n shadowing maps corresponding to n BSs covering the same area must be generated, the procedure should be as follows:

1. generate $n + 1$ matrices, $g_0, g_1, \ldots, g_n$, every element of which is a Gaussian random variable with zero mean and standard deviation σ;
2. given a correlation coefficient of shadowing from different BSs equal to ρ, produce n shadowing maps, according to

$$G_i = \rho^{1/2} g_0 + (1 - \rho)^{1/2} g_i, \quad i = 1, 2, \ldots, n \quad (4.7.9)$$

3. compute the two-dimensional inverse Fourier transform to obtain $h(x, y)$;
4. use two-dimensional convolution to filter each shadowing map, G_i, to obtain the desired filtered maps.

This scheme for simulating shadowing provides the basis for carrying out more realistic system level simulations, since it provides a link between MT location and shadowing, which is coherent with the environment description.

4.7.5.5 Narrowband short-term fading

The rapid variations of the received signal over distances of the order of a few wavelengths, or short periods of time, due to the multipath channel behaviour, i.e. the changes in the magnitude and phase of arriving waves, is referred to as small-scale or short-term fading. Globally, the received signal magnitude is described by the superimposition of the distance-dependent path-loss value, long- and short-term fading effects.

Short-term fading results from the path length difference between waves arriving from different paths, i.e. from scatterers located within the propagation environment. These differences lead to significant phase differences.

In the narrowband case, all rays arrive essentially at the same time when compared to the system resolution; hence, the short-term variations of the magnitude of the received signal can be represented by the sum of an LoS component (if it exists) plus several reflected and/or diffracted ones, being usually described by a Rayleigh or Rice distribution, depending on the existence of a LoS component [Pars92].

Under LoS (Rice case), the ratio between the power of the LoS component and the power of the reflected/diffracted ones is usually named the Rice or Ricean factor, K. It should be noted that the value of K is a useful measure of the communications link performance, its proper estimation being of practical importance for an accurate channel characterisation [Corr01].

When the long-term fading effect is modelled by a log-normal distribution, and the Rayleigh one is used to model the short-term fading, the overall signal distribution can be modelled by the Suzuki distribution [Pars92].

It is usually well accepted that long- and short-term effects can be modelled separately. In [Ande02], it is shown that the long-term fading in general may not be due to shadowing, but rather to the slow variation of the coupling between scatterers, when the MT is moving. This means that the long-term fading is just as unpredictable as the short-term one, since it originates from the same scatterers. Shadowing will still exist behind major changes in the environment.

A new model using multiple scattering, replacing the traditional subdivision of the total fading into a slow log-normal and a fast Rayleigh component is proposed in [Ande02]. The physical basis is a model of forward scattering between scatterers, introducing multiple scattered waves defining a new transfer function, H. This function consists of a sum of a small number of terms, where each term is a multiple product of complex Gaussian distributions

$$H = K + H_1 + \alpha H_2 H_3 + \beta H_4 H_5 H_6 + \cdots \quad (4.7.10)$$

where H_i are complex, independent Gaussian variables (Rayleigh in magnitude), and K is a constant related to the Rice factor. Parameters K, α and β are found by minimising the mean square error against measurements.

The double fading has been described earlier by [EFLR97] as cascaded Rayleigh fading, and it was shown that the probability density for the power, s, is given by

$$p(s) = 2K_0(2\sqrt{s}) \quad (4.7.11)$$

where K_0 is the modified Bessel function of the second kind and zero order. The Cumulative Distribution Function (CDF) is then given by

$$\text{Prob}(s' \leq s) = 1 - 2\sqrt{s}K_1(2\sqrt{s}) \quad (4.7.12)$$

with K_1 representing the modified Bessel function of the second kind and first order.

The advantage of this distribution is the insight it gives into the origin of the long-term fading, while its

disadvantage is the lack of a simple analytical function for the PDF of the received power except in special cases, e.g. the double product of Rayleigh fading paths.

This new model can be made similar in shape to the Suzuki distribution, but it has a different interpretation. The log-normal (refer to Suzuki) distribution is usually interpreted as a shadowing function, which influences the local mean value, the shadowing being dependent on the local environment. The proposed distribution has a constant mean power for the single scattering for the whole environment, and the variation of the mean of the total power stems from the slowly varying scattering between the scatterers as the MT moves. Thus, there is no need for a shadowing argument to explain the short-term fading. The resulting parameters from the fitting of the distributions may be interpreted as revealing the propagation mechanisms. According to [Ande02], the long-term fading originates from the same random elements as the short-term fading.

Short-term fading can be generated in a simulation by using either RT or statistical approaches [TLVD01], [Kunn02]. In statistical modelling, the fading processes have given predefined characteristics. Moreover, statistical approaches enable simple modelling of short-term fading in both time and frequency domains. RT approaches, being site specific, are numerical intensive techniques.

Since the knowledge of the mean level of the received signal in a multipath environment, and of its variations around the mean, is fundamental when designing a wireless communications system, it should be properly assessed if different techniques are appropriate for evaluating such variations. From the work in [ThYo03], it is shown that RT techniques make it possible to estimate with sufficient accuracy not only the mean signal level but also its variations about the mean, therefore allowing to evaluate its statistics, at least in indoor environments.

The capacity to calculate the mean power and the local statistics of the received signal leads to some optimism over the possibility of obtaining site specific channel models, which can account for the channel variations in different locations inside an area of interest. Furthermore, starting from the Doppler spectrum at one single point, obtained using RT techniques, it is possible to estimate the signal statistics in a local area of that point. This possibility is of great practical importance, since it substantially reduces the local statistics calculation time.

An alternative to such techniques is the statistical modelling approach. In [Kunn02], the statistical modelling and simulation of short-term fading with Rice distributed envelope and the desired temporal, spatial and spectral correlation is considered. The proposed model applies to the simulation of MIMO multicarrier systems. The simulated short-term fading channel gains consist of a deterministic direct part plus a Gaussian one, hence having Rice distributed envelope. The Gaussian part of the channel gains can have desired temporal, spatial and spectral correlation generated by time-correlation shaping filtering and space-frequency correlation transformation. The proposed short-term fading simulator structure enables also the generation of time varying tap weights for the classical tapped delay line model.

The number of Ricean fading channel gains to be generated is given by the product between the number of subcarriers, the number of transmit antennas and the number of receive ones. Concerning the simulator structure, a noise generator outputs uncorrelated white complex Gaussian noise sequences, which are fed to parallel time correlation shaping filters that have identical impulse responses. The uncorrelated output sequences with the desired temporal correlation are fed into a space-frequency correlation transformation that is performed for each set of uncorrelated samples. The complex Gaussian samples with the desired temporal and spatio-spectral correlation are added up with complex samples representing the direct component. Finally, the sequences are interpolated to get to the desired channel sampling rate.

Recently, MIMO systems have gained significant attention due to the potential of achieving high information theoretic capacities. Central to these techniques, multiple antennas are employed at both the BS and the MT in order to exploit the multipath richness of the propagation channel. One way to achieve this is to separate the antenna elements at both the BS and the MT, so that large diversity orders can be achieved. This requires antenna spacings of up to tens of wavelengths at the BS and up to a wavelength at the MT. Employing multiple antennas at the BS does not present a significant problem, but accommodating more antennas at the MT introduces several constraints for practical implementation.

In this regard, the Electromagnetic Vector Sensor (EVS) can be deployed as a compact MIMO receiver antenna capable of measuring six time varying electric fields, and magnetic fields, at a point in space, Figure 4.7-35. Though the EVS has been widely used for direction finding applications [NePa94], [WoZo00], recent results show that EVS can be applied to mobile communication systems, whereby the use of polarisation diversity can provide capacity improvement over a conventional dual-polarised system [AnMd01].

In [ThYo03], the closed-form expression for the spatial fading correlation function in terms of azimuth and elevation angle, and geometry of the EVS as well as polarisation states of the wave, is derived. Assuming a frequency non-selective directional Rayleigh fading channel model, in a MIMO configuration, the channel impulse response matrix $\mathbf{h}(t)$ is given by

$$\mathbf{h}(t) = \sum_{m=1}^{M} \alpha_m(t) \cdot \mathbf{a}(\phi_m, \theta_m, \gamma_m, \eta_m) \qquad (4.7.13)$$

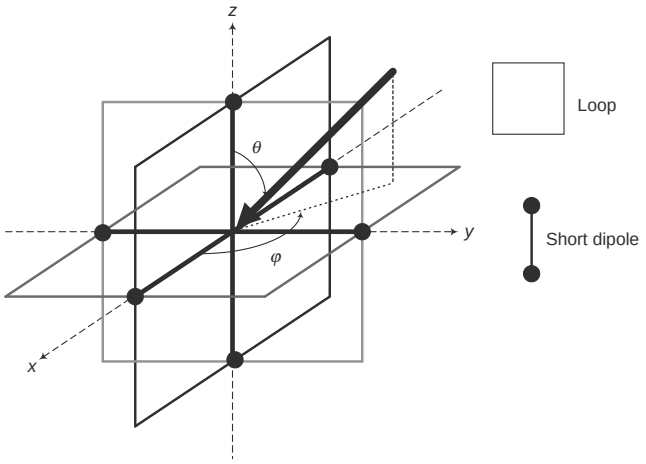


Figure 4.7-35 An EVS composed of three identical short dipoles and three magnetically identical small loops [ThYo03].

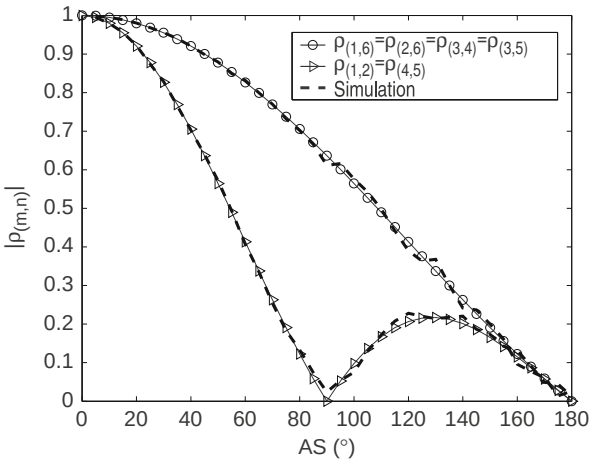


Figure 4.7-36 Spatial correlation between elements *m* and *n* of the EVS for a mean azimuth of arrival of 45° a mean elevation of arrival of 90 and an elevation spread of 0, [ThYo03].

where α_m is the complex amplitude of the mth multipath component, M is the total number of components and $\mathbf{a}(\phi_m, \theta_m, \gamma_m, \eta_m)$ is the steering vector. Scalars ϕ and θ are the azimuth and elevation angles, as illustrated in Figure 4.7-35, while scalars γ and η denote the auxiliary polarisation angle and polarisation phase difference, respectively. In [ThYo03], the angle spread in both azimuth and elevation are modelled by uniform distributions. In addition, the azimuth spread and elevation spread values are the maximum deviation of the angle spread from mean direction of arrival. Let Θ be a spatial vector parameter denoting $\Theta = [\,\phi,\, \theta,\, \gamma,\, \eta]^T$, where $[]^T$ represents transpose; therefore, the spatial fading correlation between any elements (m,n) of the EVS can be expressed as

$$\begin{aligned}\rho_{(n,m)} &= \frac{E[a_n(\Theta)a_m(\Theta)^*]}{\sqrt{E[a_n^2(\Theta)]E[a_m^2(\Theta)^*]}} \\ &= \frac{\int_\phi \int_\theta \int_\gamma \int_\eta a_n(\Theta)a_m(\Theta)^*\sin(\theta)p(\Theta)d\theta d\phi d\gamma d\eta}{\sqrt{\int_\phi \int_\theta \int_\gamma \int_\eta |a_n(\Theta)|^2\sin(\theta)d\theta d\phi d\gamma d\eta}} \\ &\quad \cdot \frac{1}{\sqrt{\int_\phi \int_\theta \int_\gamma \int_\eta |a_m(\Theta)|^2\sin(\theta)d\theta d\phi d\gamma d\eta}}\end{aligned} \tag{4.7.14}$$

where $E[]$ denotes expectation, the superscript * signifies the complex conjugate, the scalar $p(\Theta)$ is the joint PDF for all the four parameters, and a_m are the mth entries of $\mathbf{a}(\Theta)$.

Illustrative results for a mean azimuth and elevation AOA of 45° and 90°, respectively, and an elevation spread of 0°, as a function of azimuth spread, are presented in Figure 4.7-36.

4.7.5.6 Wideband short-term fading

Contrary to the narrowband case, in the wideband one the relative delays of arriving waves are large compared to the basic unit of information in the channel (symbol or bit); therefore, the wideband channel is usually modelled by a channel impulse response, where the magnitude of the different paths are taken as for the narrowband case.

In most cases, under LoS, only the first path is modelled as being Rice distributed, since this path' includes the LoS component, all other paths' magnitudes being modelled by a Rayleigh distribution. This approach was already suggested in [Corr01], where the Rice factor of the first path is derived as a function of the narrowband Rice factor, the system bandwidth, and the RMS delay spread of the propagation channel. Nevertheless, the proposed description of the Rice factor is only valid for a certain structure of the channel impulse response.

A different approach to overcome the bandwidth dependence of modelling parameters is based on the use of the time-variant transfer function instead of the time-variant impulse response, such that all known properties from narrowband models can be applied [Corr01], [Katt02].

A time-domain technique for fading depth characterisation in wideband Rayleigh and Ricean environments, described by its PDPs, is proposed in [CaCo02]. The PDF of the received power, s, is given by

$$\begin{aligned}p(s) &= \int_{-\infty}^{+\infty} \frac{2K}{a_d^2} e^{-\frac{K(2x+1)}{a_d^2}} I_0\left(\frac{2K\sqrt{2x}}{a_d}\right) \\ &\quad \times \sum_{m=2}^{M} \frac{(\lambda_m)^{M-3} e^{-\frac{s-x}{\lambda_m}}}{\prod_{\substack{k=2 \\ k\neq m}}^{M} (\lambda_m - \lambda_k)} dx\end{aligned} \tag{4.7.15}$$

where a_d is the magnitude of the LoS component, K is the Rice factor, I_0 is the Bessel function of first kind and zeroth order, and λ_m are eigenvalues of a covariance matrix obtained, by generating a matrix whose elements are given by the product between the correlation function between different frequency components, the frequency response of the filter used in the transmitting equipment, and an incremental bandwidth that depends on the system bandwidth.

When K tends to zero the Rice distribution associated to the first path degenerates to the Rayleigh one, therefore corresponding to the NLoS case.

By using this model, the fading depth for different channel models can be easily evaluated for different PDPs. When represented as a function of the Rice factor, and of the product between the system bandwidth and the RMS delay spread of the propagation channel, one gets a graph like the one in Figure 4.7-37.

The results in Figure 4.7-37 correspond to the case of a continuous exponential decaying function with decay rate equal to the inverse of the RMS delay spread, thus being representative of a large set of channel models, commonly used for simulating the propagation channel. As one can observe, for each value of K, the fading depth remains practically constant for $B\sigma_\tau < 0.02$ Hz·s. This corresponds to a situation where the system bandwidth is below the coherence bandwidth of the propagation channel, defined for a frequency correlation of 90% [Rapp86], thus signals are in a frequency flat environment. For large values of $B\sigma_\tau$, the fading depth decreases for increasing system bandwidths.

A different approach, which is based on environment properties, more precisely on the differences in propagation path length among different arriving components, is presented in [CaCo01]. The proposed approach, accounting for the maximum difference in propagation path length, Δl_{max}, and having also the Rice factor as a parameter, allows evaluating the fading depth, measured betweenp and 50% of the received power CDF, from a simple mathematical expression

$$FD_p(K, B\Delta l_{max})_{[\mathrm{dB}]} = \begin{cases} S_p(K), & B\Delta l_{max} \le \omega_{b,p} \\ \dfrac{S_p(K) - A_{1,p}(K)}{1 + A_{2,p}(K)\left[\log\left(\dfrac{B\Delta l_{max}}{\omega_{b,p}}\right)\right]^{A_{3,p}(K)}}, & B\Delta l_{max} > \omega_{b,p} \end{cases} \quad (4.7.16)$$

where $S_p(K)$, $A_{1,p}(K)$, $A_{2,p}(K)$ and $A_{3,p}(K)$ are mathematical functions of K that depend on p, and $\omega_{b,p}$ is a breakpoint value depending also on p.

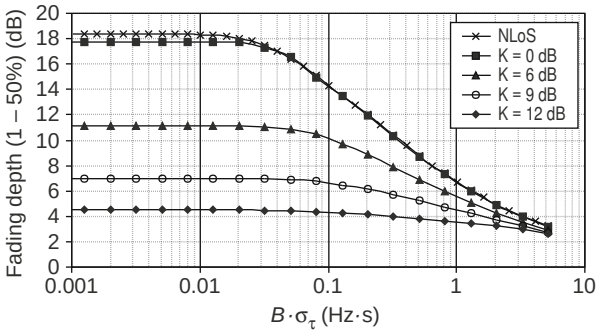


Figure 4.7-37 Fading depth, exponential model [CaCo02].

For any given value of p, functions $S_p(K)$, $A_{1,p}(K)$, $A_{2,p}(K)$ and $A_{3,p}(K)$ can be evaluated from

$$S_{[\mathrm{dB}]} = \frac{(b_1 - b_2)}{1 + b_3\left(\frac{K_{[\mathrm{dB}]}}{10} - b_4\right)^{b_5}} + b_2 \quad (4.7.17)$$

$$A_{1[\mathrm{dB}]} = c_{11}\arctan(c_{12}K_{[\mathrm{dB}]} - c_{13}) - c_{14} \quad (4.7.18)$$

$$A_2 = c_{21}\left[\frac{\pi}{2} - \arctan(c_{22}K_{[\mathrm{dB}]} - c_{23})\right] + c_{24} \quad (4.7.19)$$

$$A_3 = c_{31}\arctan(c_{32}K_{[\mathrm{dB}]} - c_{33}) + c_{34} \quad (4.7.20)$$

where parameters b_i ($i = 1, \ldots, 5$) and c_{jk} ($j = 1,\ldots, 3; k = 1, \ldots, 4$) depend on the value of p. Values for the parameters b_i and c_{jk}, for $p = 10\%$, are shown in Table 4.7-9. Values for different values of p can be found in [CaCo01] and [Card04].

Since there is a close relation between the physical and geometrical environment characteristics and the PDP of the propagation channel (hence, the RMS delay spread), a simple relationship between the maximum difference in propagation path length among different arriving components and the RMS delay spread of the propagation channel can be established, therefore bridging the gap between the two previous models [CaCo03]

$$\Delta l_{max} = c\sigma_\tau\frac{K+1}{\sqrt{K}} \quad (4.7.21)$$

where c is the speed of light.

Table 4.7-9 Parameters b_i and c_{jk} for p = 10%.

	S		A_1	A_2	A_3
b_1	8.080	c_{j1}	0.289	0.141	0.452
b_2	0.070	c_{j2}	0.225	0.338	0.180
b_3	0.690	c_{j3}	0.349	2.650	0.689
b_4	−0.410	c_{j4}	0.421	0.610	2.295
b_5	2.943			–	

This allows one to use any of the proposed approaches for evaluating the fading depth in a given environment, defined either by physical and geometrical properties or by the PDP. This approach, being simple, is effective for evaluating the fading depth in different environments and for different system bandwidths, while allowing to use any of the described approaches, starting either from the environment characteristics or from the PDP of the propagation channel.

With the emergence of third generation systems, improvements in digital signal processing hardware, and the increasing demand for a larger capacity, the spatial domain appears to be one of the last frontiers for exploiting the possibility of increasing systems capacity. This can be achieved by using different types of antenna, at either the BS, or the MT, or both; therefore, new antenna and signal processing techniques are emerging, including smart antennas, either adaptive or switched beam, spatial diversity combining, and MIMO.

A lot of work has been done concerning the implementation of algorithms for achieving the desired link quality and system capacity using these techniques; however, it is not clear how the values of short-term fading depth observed by the different systems, working in different environments, depend on the antenna arrays being considered, namely on their half-power beamwidth. In [CaCo04], the approach in [CaCo01] is extended, by including the influence of the antenna's radiation pattern. This influence is modelled through the variation of the Rice factor and the maximum possible difference in propagation path length among different arriving components, relative to the case of using omnidirectional antennas.

The variation of the Rice factor, as a function of the antenna half-power beamwidth, $\Delta K(\alpha_{3\text{dB}})$ depends on the statistical distribution of AoAs (or AoDs) and on the type of antenna being considered. Assuming an ideal directional antenna (hypothetical antenna with constant gain within the antenna half-power beamwidth and zero outside) and a truncated Gaussian distribution of AoAs (or AoDs) with standard deviation, σ_s, the value of $\Delta K(\alpha_{3\text{dB}})$ is evaluated as

$$\Delta K(\alpha_{3\text{dB}})_{[\text{dB}]} = 10\log\left[\frac{1-2Q(\frac{\pi}{\sigma_s})}{1-2Q(\frac{\alpha_{3\text{dB}}}{2\sigma_s})}\right] \qquad (4.7.22)$$

where $Q(\cdot)$ is the well-known Q-function [Carl86].

The variation of the maximum possible difference in propagation path length among different arriving components depends on the scattering model being considered, e.g. Geometry-based Stochastic Channel Models (GSCMs) are usually used, elliptical and circular ones being the most common for simulating micro-, pico- and macrocellular environments. The results in [CaCo04] illustrate such approach.

In the literature on measurements and modelling of time-variant radio channels, there has almost always been a confinement to moving transmitters or receivers in a static environment. For applications of practical interest so far, like mobile cellular phones, this was obviously realistic. Although moving scatterers are certainly always present, their influence is negligible, at least in outdoor environments, due to their scattering cross-section being small compared with fixed scatterers [Cox73].

For indoor environments, the scattering cross-section of moving objects (persons, doors, windows, etc.) still cannot be neglected compared with fixed scattering objects (e.g. walls and furniture). Furthermore, for some applications, e.g. wireless LANs, MTs are usually fixed during operation, whereas quite often some scatterers are moving in the environment [KaFr01].

Some of the few investigations of the influence of moving scatterers that can be found in the literature are related to the temporal power fluctuations of a narrowband channel, e.g. [HMVT94]. For a more comprehensive and more universal treatment, however, wideband measurements are necessary. For this purpose, wideband measurements were performed in a laboratory room with a distance of 6.6 m between transmitting and receiving antenna [KaFr01]. The LoS path was partially obstructed

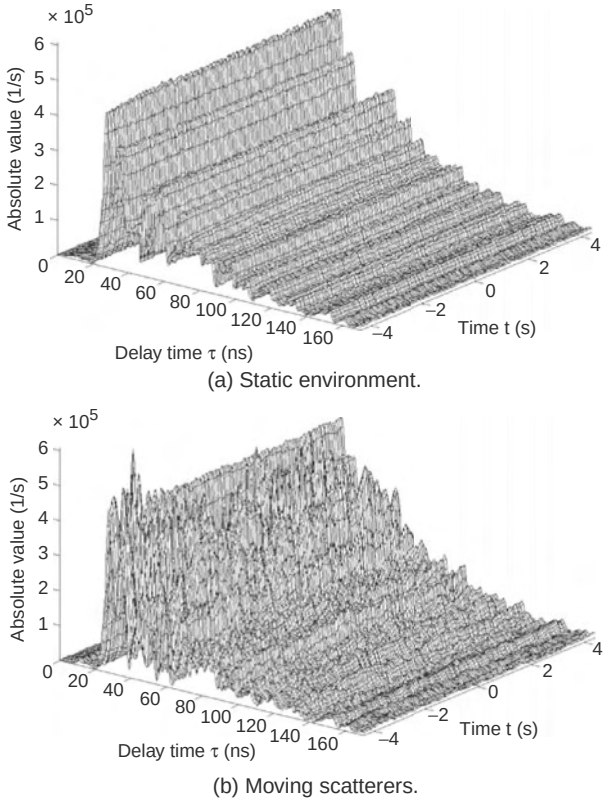


Figure 4.7-38 Time-variant impulse response at 1.8 GHz with fixed antennas [KaFr01].

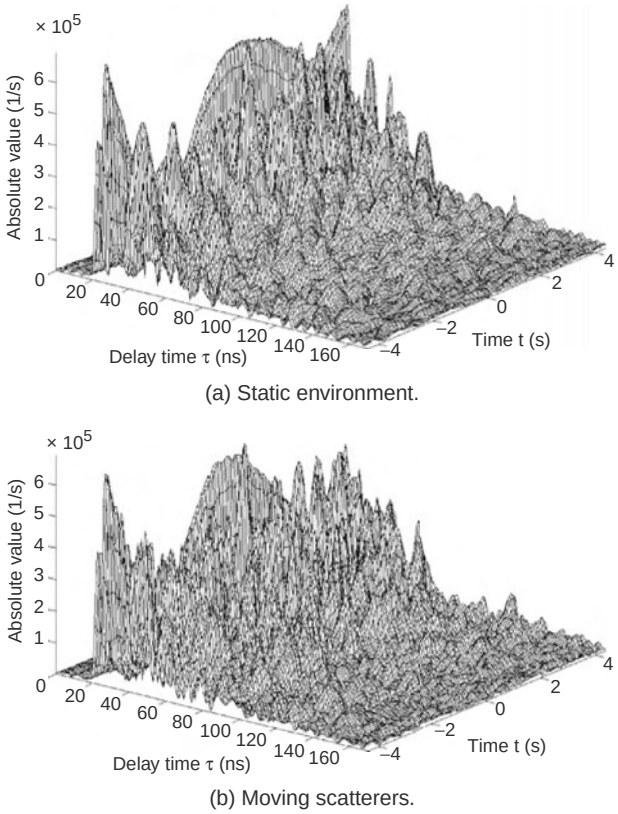


Figure 4.7-39 Time-variant impulse response at 1.8 GHz with antennas moving [KaFr01].

by furniture and laboratory equipment. For most of the measurements, three persons acted as moving scatterers.

Systematic measurements at 1.8, 5.2 and 17.2 GHz have been performed with a bandwidth of 600 MHz. The results from measurements have shown that, for the case of moving antennas, there is a fundamentally different fading behaviour and also a fundamentally different shape of the time-variant impulse response. This fundamental difference leads to the conclusion that it is not allowed (though quite often done) to uncritically adopt results and models found from investigations of channels with linearly moving antennas to applications with moving scatterers in an otherwise static environment, Figures 4.7-38 and 4.7-39.

Globally, these results have shown that the influence of moving scatterers cannot be neglected, as it has been done quite often. Concerning the frequency dependence of the influence of moving scatterers and/or moving antennas on the channel behaviour, no fundamental differences have been found from the comparison of the three different frequency bands.

As a major conclusion, it can be stated that, for a statistical modelling, an unmodified application of the well-known models for channels with moving transmitters or receivers to channels with moving scatterers obviously seems to be not very realistic.

4.7.5.7 Temporal and angular dispersion

A number of large measurement campaigns were conducted in the context of comparison of propagation scenarios or environments or of frequency bands. Other contributions intended to cover more rare scenarios or topics. All these contributions contain data on temporal and angular dispersion of mobile radio channels. The choice is made not to remove this data from the context in which it was gathered. That leads to an organisation of this section into subsections on comparisons of environments and frequency bands, respectively, on modelling-based characterisation, on characteristics of clusters, on diffusely scattered power, on presumed static environments and on some special cases.

Comparisons between mobile radio environments

A number of COST 273 participants reported large to very large measurement campaigns that range over diverse propagation environments. This enables the study of the similarities and differences between environments. Salous and Hinostroza conducted extensive measurement campaigns in seven indoor and outdoor-to-indoor environments in Manchester, UK, including a gymnasium, a laboratory, and floor-to-floor and building-to-building links [SaHi03]. The measurements were done at 2.35 GHz in 300 MHz bandwidth with the proprietary UMIST chirp sounder; overall dynamics were stated to be better than 40 dB. From the measurements, path loss, power delay profiles, delay spread, and coherence bandwidth were determined. The coherence bandwidth ranged from 20 to 250 MHz, the RMS delay spread from 25 to 150 ns for a maximum excess delay of 800 ns. For rooms, halls, and floor-to-floor links, Ricean distributions fitted the delay spread distribution best, for LoS in-building corridors log-normal distributions gave the best fit. The authors noticed a distinct dependence of the delay spread on the clipping level: considering only multipath components stronger than −20 dB under the instantaneous maximum amplitude gave a median in-building delay spread of around 27 ns while for an instantaneous dynamic range of 35 dB this figure became 44 ns [SaHi03].

For an outdoor-to-indoor setting at 5.2 GHz in Lund, Sweden, Wyne *et al.* reported smaller dispersion figures [WAKE04], although this could well be contributed to the geometry of the measurement set-up and not to the different frequency band. Median values for the RMS delay spread were 7–10 ns for rooms facing the TXs, 12–14 ns for rooms at the opposite side of the wing, and 8–12 ns in between, in the corridor. Spreads on AoA were almost independent of receiver or transmitter location (about 0.24π or 14°) but the spreads of the AoD

Table 4.7-10 Angular spreads for various measurement environments (after [KKVV05]).

Environment	RX azim. spread [°]	TX azim. spread [°]	RX elev. spread [°]	RX elev. spread [°]
Indoor	37.3	40.0	7.0	6.3
Outdoor–indoor	39.2	7.0	5.8	2.2
Microcell (LoS)	28.9	5.1	2.5	1.3
Microcell (NLoS)	40.3	12.6	4.7	2.5
Macrocell	52.3	7.6	7.7	1.7

clearly depended on the TX position, ranging from 0.03 to 0.08π or 1.6 to 4.4° [WAKE04].

A comparison between spreads at 5.3 GHz in different environments was made by Kolmonen *et al.* [KKVV05]. They held an extensive measurement campaign in typical indoor, outdoor-to-indoor, and micro- and macrocellular urban environments (downtown Helsinki). For the microcellular environments, the TX antenna was positioned well above surrounding rooftops. Dual polarised antenna arrays were used and directional information was retrieved separately at the RX and TX side by beamforming, as a fast and robust method. From that, angular spreads, both in elevation and azimuth, were calculated according to 3GGP TR25.966. First results on the angular spreads in the respective environments are given in Table 4.7-10, showing smaller angular spreads at the TX side than at the RX side, except for the indoor environment where both spreads are comparable in value. The influence of the differences in the antennas' operational angular ranges was not discussed.

In grossly the same environments, Kainulainen *et al.* made a large polarimetric survey at 5.3 GHz, determining Cross-polarisation Power Ratios (XPRs) and cross-correlation coefficients between co-polar and cross-polar links. The XPRs are shown to be approximately log-normally distributed in all environments, but with different means and standard deviations, Table 4.7-11. The cross-correlation coefficients, at both the transmit and receive side, turned out to be low, between 0.26 and 0.45 for NLoS links. In the LoS microcellular environment, the range was 0.51 to 0.62, Table 4.7-12.

Table 4.7-11 Mean XPRs and standard deviations in diverse environments at 5.3 GHz (after [KaVV05]).

Environment	$\overline{XPR_V}$ [dB]	$^{\sigma}XPR_V$ [dB]	$\overline{XPR_H}$ [dB]	$^{\sigma}XPR_H$ [dB]
Indoor picocell 1	4.5	2.3	5.8	2.9
Indoor picocell 2	4.9	2.3	5.7	2.4
Outdoor–indoor	6.2	3.9	7.3	3.3
Microcell (LoS)	8.6	1.8	9.5	2.3
Microcell (NLoS)	8.0	1.8	6.9	2.8
Macrocell (NLoS)	7.6	3.4	2.3	3.1

The differences between microcellular and macrocellular scenarios in the same area were the subject of an investigation by Hugl *et al.* [HuKL02]. They changed BS antenna heights in order to study the influence on the mean AoA for vehicular channels in a densely built-up urban environment, downtown Helsinki with six- to seven-storey buildings. The BS antenna height was 10 m above rooftop level or around rooftop level, reflecting either macro- or microcellular coverage. The 20 km measurement route covered all accessible streets within a 120° sector from the BS up to about 1.5 km distance. The authors used a Bartlett beamformer for determining the dominant angles of arrival and compared these with the geometrical AoA, the direction of the straight line between TX and RX. For the macrocellular antenna placement, the dominant and geometrical AoA matched quite well, Figure 4.7-40. However, for the microcellular set-up, large differences were found. The authors assume this to be caused by the difference between over-rooftop propagation and propagation through street canyons. The down-tilting of the receiving antenna could be an additional factor in excluding over-rooftop components at the lower BS position, due to the narrow elevational pattern.

Frequency-band dependent dispersion

For present system bandwidths, statistical channel properties are generally assumed not to vary over the system bandwidth. For operation over different bands, for instance UMTS FDD operation or Multi-Band OFDM (MB-OFDM), and for UWB applications, such invariability is not obvious. Jämsä *et al.* measured delay spreads in a total bandwidth of 10 GHz, centred around 6 GHz, and compared some 500 MHz wide subband results in order to study the frequency dependence of indoor propagation [JäHH04]. Also, a comparison was

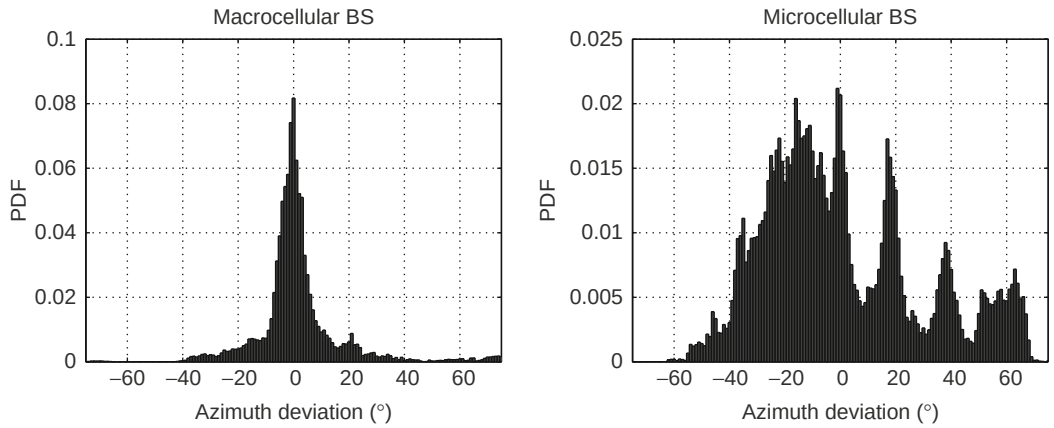
Table 4.7-12 Mean cross-polarisation cross-correlation coefficients in diverse environments at 5.3 GHz (after [KaVV05]).

Environment	Transmit correlation		Receive correlation	
	$\overline{\rho(h_{VV}, h_{VH})}$	$\overline{\rho(h_{HH}, h_{HV})}$	$\overline{\rho(h_{VV}, h_{HV})}$	$\overline{\rho(h_{HH}, h_{VH})}$
Indoor picocell 1	0.30	0.31	0.35	0.30
Indoor picocell 2	0.26	0.28	0.29	0.32
Outdoor–indoor	0.35	0.37	0.37	0.45
Microcell (LoS)	0.52	0.56	0.51	0.62
Microcell (NLoS)	0.39	0.40	0.37	0.45
Macrocell (NLoS)	0.30	0.33	0.35	0.31

made between results calculated from 500 MHz-wide cuts from the 10 GHz-wide measurements and from 500 MHz that were individually measured at maximum frequency resolution, rendering 1601 frequency points instead of 81. The full resolution sweeps showed higher dynamic range, but the authors conclude nevertheless that UWB measurements can be used for narrowband modelling with additional benefits from using the same data for different subbands. On the frequency dependence, they report that no real trends with frequency had been seen, apart from the strong dependence of signal strength with the frequency band. Although, in general, delay spread tends to decrease with increasing frequency, Table 4.7-13

Spatial correlation behaviour over a similar bandwidth was studied by Liu *et al.*, in this case over 7.5 GHz bandwidth centred around 6.85 GHz. Their aim was to determine practical antenna separations for MB-OFDM applications [LAME05]. The measurements were taken in a single workshop room of $6 \times 6\,\mathrm{m}^2$. The distance at which the spatial correlation coefficient was lower than 0.6 sank from 2.2 cm at 4 GHz to 0.4 cm at 10 GHz, roughly inversely proportional with frequency. The differences between LoS and NLoS, the latter achieved by placing a large grounded aluminium sheet between TX and RX, are minor as correlation distance is concerned.

Bultitude and Schenk measured NLoS vehicular metropolitan channels at 1.9 and 5.8 GHz in a Manhattan grid, in downtown Ottawa [Bult02a], [SBAP02]. The measurements were taken from the top of a van in 10 MHz bandwidth and the measurement runs were up to 1 km in length. The data analysis was done per street, in order to be able to separate parallel from perpendicular streets, all NLoS. Bultitude and Schenk found consistently lower delay spreads at 5.8 GHz (16% lower) and larger coherence bandwidths. The median delay spreads were 213 ns and 184 ns at 1.9 GHz and 5.8 GHz, respectively. The corresponding single-sided coherence bandwidths were around 1.8 MHz and around 1.9 MHz. They attribute this partly to higher diffraction losses at higher frequencies. Fresnel zones are smaller too at higher frequencies but it is unclear whether this results in



(a) Macrocellular case. (b) Microcellular case.

Figure 4.7-40 Deviation between geometrical angle and the dominant AoA (after [HuKL02]).

Table 4.7-13 Measured indoor RMS delay spread for different frequency bands, NLoS and LoS in an empty room. UWB results are subband results cut from a full 10 GHz sweep, WB results are swept individually over 500 MHz (after [JäHH04]).

Centre freq. [GHz]	Band-width [MHz]	RMS delay spread LoS/[ns]UWB	RMS delay spread NLoS/[ns]UWB	RMS delay spread LoS/[ns]WB	RMS delay spread NLoS/[ns]WB
1.25	500	6.3	8.4	10.9	11.3
5.25	500	4.5	7.4	10.4	7.3
10.75	500	3.4	5.9	8.6	–
6	10,000	6.4	6.2	–	–

better propagation conditions for multipath components or in fewer available scatterers [Bult02a].

Differences between UMTSFDD duplex bands in urban environments were investigated by Salous and Gokalp, for Manchester city centre [SaGo01], and by Foo *et al.* for Bristol city centre and suburbs [FBKE02]. Salous and Gokalp used the UMIST proprietary channel sounder, to measure pedestrian and vehicular channels at two 60 MHz bands simultaneously, centred at 1950 and 2140 MHz, the UMTS uplink and downlink band respectively [SaGo01]. The measurement results were compared between these two bands and between 5 MHz subbands within the 60 MHz bands. The authors note that frequency dependent differences between subbands or bands are determined by the amount of dispersion on the channel. For the delay spread results, larger variations between subbands were seen in NLoS than in LoS situations and larger differences between uplink and downlink were found in environments with dense scattering than with sparse scattering. It was noted that locally, frequency differences between subbands or bands are larger than apparent from CDFs over pooled data. These differences were largest at largest range or smallest link budget, presumably because dispersion is strongest in these cases [SaGo01].

The latter conclusion was reached by Foo *et al.* too [FBKE02]. In this case, two 20 MHz bands centred at 1920 and 2120 MHz respectively, were measured simultaneously from a moving vehicle. Time of arrival, AoA, and power of multipath components were estimated using a 2D unitary ESPRIT algorithm. Different responses were seen, both temporal and angular, when comparing both bands. The largest difference, between two environments and between both bands, was found for the (azimuthal) angular spread, with strong decorrelation between the two frequency bands. The decorrelation is strongest in urban environment, which has a richer scattering than the suburbs, as measured from the larger angular spreads. The authors conclude from their findings that downlink beamforming based on uplink measurements in scattering-rich environments is sub-optimal [FBKE02].

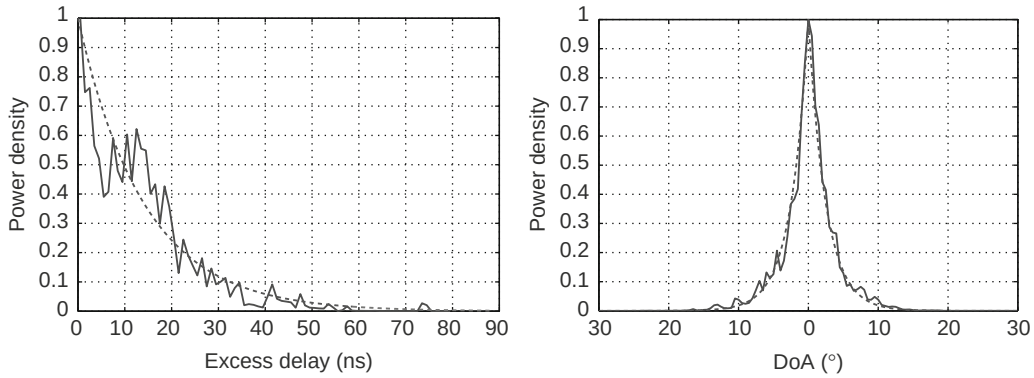
Modelling-based characterisation

The common approach to characterisation of mobile radio environments is by measurements. Nevertheless, for simple user scenarios and simply structured environments, analytical modelling can still provide insight in fundamental phenomena. Hansen calculated power delay profiles for chambers of a cubic shape for modelling short-range communication as in WLAN or ad-hoc networking applications [Hans03]. Based on ray-optical theory and under the assumptions of well-mixed polarisations, uniformly distributed directions of travels, and random phases of scattered components, he arrives at a PDP of the shape of $\exp(-\tau/\sigma_\tau)$ for cubic rooms, in which

$$\sigma_\tau = (2d\xi_\tau)^{-1} \tag{4.7.23}$$

In turn, $d = -ln(d_E)$, with d_E the mean reflection attenuation, and ξ_τ represents the room dimensions in terms of number of wall interactions per unit of time. As for exponential decay in general, σ_τ equals the delay spread. Hansen claims that delay spread is a robust parameter with respect to modelling inaccuracies and that therefore the derived model will be a good approximation to many propagation scenarios. He compared his model with measurements, at 5.25 GHz with 100MHz bandwidth, in rooms and a corridor in an office environment. The measured RMS delay spread values were small, ranging from 7 to 13 ns. The analytically derived values could well serve as upper bound as these values generally exceeded the measured ones with one standard deviation or more [Hans03].

In contrast to the approach of Hansen, Fügen *et al.* turned to advanced simulation tools as most of the realistic environments are too complex to use analytical channel models [FMDW03]. They used the proprietary 3D RT tool of Karlsruhe University in an urban macrocellular environment without high-rise buildings, Karlsruhe, over an area of approx. 500 × 700 m^2. The simulation incorporates multiple reflections, multiple diffractions, and combinations of both. The BS was virtually placed at 25.5 m height, well above the average



(a) Digital power spectral density. (b) Azimuthal power spectral density.

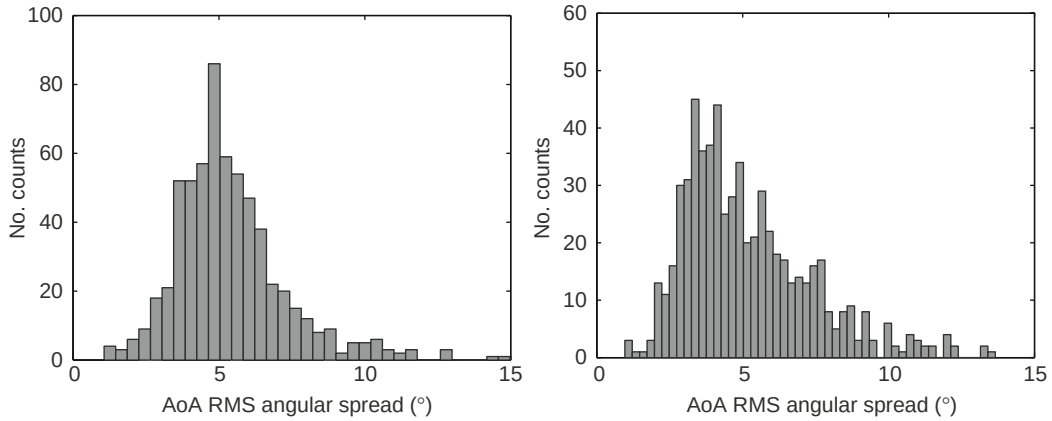
Figure 4.7-41 Intracluster marginal power density distributions for LoS, measured (solid line) vs. modelled (dashed) (after [CTLM02]).

building height of 12.5 m and the carrier frequency was 2 GHz. The results show that the angular spreads at the MT are higher on street crossings than in street canyons. Also, the authors model both delay and angular spread, at BS and MT, with an offset negative exponential function of distance between BS and MT. A disadvantage is that the modelled delay spread falls with distance.

Intracluster dispersion

The concept of clusters of scatterers has been brought into radio channel modelling, both to ease and to refine the statistical description of mobile channel behaviour The scattering environment is in this concept seen as consisting of a distributed set of concentrations of scatterers, the clusters. Description of the properties of these clusters of scattering objects was undertaken by a number of groups within COST 273. Chong *et al.* took measurements at 5.2 GHz with 120 MHz bandwidth in a large office and a large open space, during office hours [CTLM02]. An FD-SAGE for estimation of multipath components was used and the delay-azimuth estimates were clustered by a 2D Gaussian kernel density estimator; a cluster was defined as a set of multipath components with similar excess delays and AoAs. Chong *et al.* propose a cluster model, to be discussed in the chapter on MIMO modelling, in which the temporal and spatial domain are independent for Obstructed Line-of-Sight (OLoS) and NLoS links, but are coupled for LoS. For LoS, the intracluster marginal power density distributions showed exponential decay with excess delay and were Laplacian for the azimuthal power density, Figure 4.7-41.

An alternative method by Czink and Yin uses a Bartlett beamformer to coarsely determine clusters in the AoA-AoD domain. SAGE estimates of corresponding AoAs and AoDs derived from instantaneous channel realisations were then selected with this cluster information [CzYi05]. From the clustered estimates, angular spreads were calculated. On synthetic data, spread estimates were shown to be approximately unbiased. The authors observe on measured data different intracluster angular spreads at the TX and RX side [CBYF05], likely due to the position of the TX antenna in the corridor, Figure 4.7-42. This method is an improvement of an



(a) AoA RMS angular spread. (b) AoD RMS angular spread.

Figure 4.7-42 Histograms of cluster angular spread estimates for measured indoor data (after [CBYF05]).

Table 4.7-14 UWB cluster properties in residential environment (after [HaTK05]).

Scenario	No. clusters	No. paths/ cluster	Delay spread [ns]	Angular spread [°]	Fraction power [%]
Room (LoS)	7	3–24	0.95–1.38	1.9–4.5	73.4
Room–corridor (OLoS)	6	5–23	0.54–1.51	2.4–6.7	50.2
Room–corridor (NLoS)	6	3–27	0.64–1.65	0.9–6.4	50.2
Inter–room (NLoS)	8	6–20	0.21–1.28	1.8–6.3	66.8
Inter–floor (NLoS)	10	3–25	0.07–1.28	3.0–18.2	50.6
Indoor–outdoor (OLoS)	3	15–40	0.25–2.30	3.1–7.6	73.8
Indoor–outdoor (NLoS)	3	21–26	0.10–2.07	2.7–7.1	81.3

earlier attempt, based on AoA-AoD estimations on eigenmode channels [CHÖB04], that consistently underestimated cluster spreads.

In the UWB band from 3.1 to 10.6 GHz, Haneda *et al.* also determined cluster properties, in a wooden Japanese residential house [HaTK05]. In total 100 paths were extracted, using a SAGE algorithm, that were clustered heuristically in the delay-AoA (azimuth) domain. The results are given in Table 4.7-14 and are lower than the ones documented above, probably due to the short paths, the light construction and the relatively small size of the building. The entries in the last column of Table 4.7-14 refer to the amount of power that could be recovered by the SAGE algorithm, indicating that roughly a quarter to half the power could not be modelled, see the next section.

Diffusely scattered power

The question of the amount of diffusely scattered power in mobile propagation environments is still largely unanswered. This is of importance for high-resolution parameter estimation applications, where generally a discrete scatterer model is assumed, and from the point of view of MIMO capacity. Degli-Esposti *et al.* approached the problem from the modelling side by examining to which extent diffusely scattered components determine delay and angular spread in urban environments [DEESGK02b]. They did this by comparing 3D RT results with measurements. The measurements were made in the centre of Helsinki, along a waterfront and amid high-rise buildings, at 2154 MHz. The RT algorithm used is based on a full 3D image RT technique [DDEFR] with diffuse scattering incorporated according to the effective roughness approach [Degl01], [DEFVG05]. The prediction of azimuthal angular spread is reported to be good in general. Delay spread was predicted fairly well too; Figure 4.7-43 shows the results for a quasi-LoS route. In this case, the authors concluded that modelling of diffuse scattering from far-away buildings across bays helps to reconstruct the delay spread distribution over the measurement route. But, for about 10% of the routes the model does not perform well, especially when the BS is above rooftops and the mobile sunk into a high-rise urban structure. Degli-Esposti *et al.* assume that in these cases the building database could be inaccurate (see Section 4.7.4.3) and/or that not enough buildings are considered in the RT so contributions from distant scatterers are missing [DEESGK02b].

An approach of analysing large sets of measurements was taken by Trautwein et al. [TLST05]. They quantified the amount of diffusely scattered power using the Rimax estimation algorithm pioneered by Richter and Thomä that jointly estimates AoA, AoD, delay, Doppler, and polarisations [TLRT05a], see also Section 4.7.4.4. After subtracting the estimated coherent information from the channel transfer, the algorithm parametrises the residual, using the delay domain, into noise and an exponentially decaying diffuse power burst. This power burst is assumed to comprise all Diffuse Multipath Components (DMC). Large differences between environments were found: in an outdoor small macrocell the median value of the ratio between power in specular and in diffuse components was 14 dB while in the large main railway hall

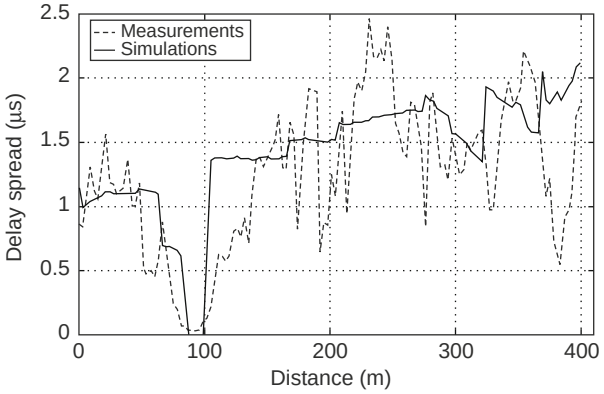


Figure 4.7-43 Comparison between measured delay spread and that simulated by RT (after [DEESGK02b]).

Table 4.7-15 No. of paths and ratio of power in specular and diffuse components for various environments (after [TLST05]).

	No. paths		Power ratio specular/DMC				
Environment	Mean [dB]	σ [dB]	Mean [dB]	σ [dB]	10% [dB]	50% [dB]	90% [dB]
Small urban macrocell	44.4	5.4	13.1	3.4	8.1	14.0	17.0
Urban microcell	20.3	5.5	3.6	4.2	−2.1	3.7	9.0
Urban microcell	17.1	6.0	0.9	4.1	−4.1	0.4	6.8
Urban microcell, square	19.5	4.8	−0.3	5.0	−6.1	−1.2	7.8
Urban microcell, square	27.1	5.0	7.7	3.0	3.6	7.7	11.6
Urban microcell, square	30.2	6.1	2.1	3.8	−2.1	1.7	7.6
Indoor microcell lobby	26.3	7.1	0.8	3.6	−4.6	1.3	5.2
Hotspot railway station	20.4	4.9	−11.3	4.5	−19.0	−10.3	−6.5
Hotspot railway station	21.9	7.4	−9.2	6.9	−19.2	−9.0	0.4
Picocell highway	20.5	6.1	8.5	5.2	1.1	9.2	14.5

with metal support structures in München this ratio was −10.3 dB, Table 4.7-15. Other environments, such as urban microcell streets and squares, an indoor auditorium lobby, and a highway bridge picocell, showed power ratios in between. In the environments with relatively large amounts of DMC power, the estimator tended to find fewer specular components than in environments with relatively high power in the specular components.

The observation by Trautwein *et al.* that diffusely scattered power cannot be neglected in a number of environments, was made by others within COST 273 too, i.e. they experienced that there seems to be a ceiling as regards the amount of power that can be captured by high-resolution estimators. Wyne *et al.* retrieved in an outdoor-to-indoor environment a worst case of 60% with 40 multipath components, with a median around 85% and the most probable extraction efficiency of around 90% [WAKE04]. Haneda *et al.* found slightly lower figures indoors, between 50 and 80%, Table 4.7-14, although they forced 100 specular components to be estimated, despite the high temporal resolution [HaTK05].

Static channels

The influence of moving scatterers on the channel, as compared to moving a receiving antenna, was investigated by Kattenbach and Früchting [KaFr01]. To this end, they made measurements at three different frequencies, at 1.8, 5.2 and 17.2 GHz respectively, in 600 MHz of bandwidth for good temporal resolution, using a proprietary stepped-frequency correlation sounder. People walking around in the room acted as moving scatterers. The effect on static channels was clear from the broadening of the otherwise very narrow Doppler spectrum and the changes in the impulse response. Also, the amplitude variations over time are better noticeable than phase fluctuations, as many multipath components, stemming from for instance the constructional part of the building, do not change at all. The effect of moving scatterers was also distinctly different from time-variant channel behaviour due to receiver movement. Therefore, Kattenbach and Früchting conclude that channels with moving scatterers should be modelled differently from those only showing time variance due to receiver movement. For modelling temporal correlations, the moving scatterers can be ignored irrespective of their number, but they cannot be ignored for modelling amplitude and phase variations. The authors found no fundamental frequency dependence over the three measurement bands.

Medbo, Berg, and Harrysson specifically measured and modelled temporal variations in a NLoS WLAN scenario without receiver movement [MeBH04]. Overnight, for 15 hours, measurements were taken at 5.25 GHz in 200 MHz bandwidth with a VNA on a 5 s time sweep. The variations seen are believed to be caused by persons

moving in the vicinity of the antennas at either end. The terminal was placed in a room next to a corridor, the BS/access point in the corridor. They characterised the measured transfer variations by a deviation parameter p: the total power, as function of time interval, in the deviation from the mean measured transfer spectrum, the mean determined over intervals of 1000 s, i.e. 100 subsequent measurements. The total received power itself showed to be very constant over the 15-hour period. Without persons around during night time, p was of the order of −50 dB relative to the total received power; during daytime on rare occasions it was higher than −10 dB, Figure 4.7-44. The authors recommend to use p as an indicator for non-static measurement situations in for instance MIMO measurements [MeBH04]. Two different causes of channel dynamics were identified, disturbances with low Doppler shifts and those with high Doppler shifts, with the first most likely related to people standing or sitting close to the antennas, at either side, and the second by people walking by. Medbo, Berg, and Harrysson modelled the variations as the effect of moving persons, consisting of a cylindric scattering and a shadowing component simplified as knife-edge diffraction. Qualitatively good correspondence between measurements and model was reported [MeBH04].

Skentos *et al.* examined static rooftop-to-street scenarios for a wide, busy avenue in Athens under LoS propagation conditions [SKPC04]. The receiver was placed on top of a building at 10 m height, the transmitter at four positions at street level, 1.9 m above ground, between 115 to 236 m away from the transmitter. The carrier frequency was 5.2 GHz, the bandwidth 120 MHz and eight-element uniform linear arrays were used at both ends. The data was tested for WSS by means of a power correlation metric on the power delay profiles, averaged over all 64 SISO links. The threshold for the correlation coefficient with respect to measurement time was arbitrarily set at 0.75; all data passed. Typical values for temporal dispersion were 25 ns mean excess delay and 41 ns RMS delay spread. The measured channel was slowly varying with time, with the maximum observed Doppler shift around 1.5 Hz. The angular dispersion was smaller for the rooftop-positioned RX than for the TX at street level. The clustering of multipath components in the azimuth-delay plane was evident, both for RX and TX, but no cluster parameters were extracted [SKPC04]. Skentos *et al.* also investigated links from rooftop to rooftop as a typical fixed wireless access scenario. In this case, dispersion was found to be very slow, the maximum Doppler shift reported was 0.63 Hz, at 5.2 GHz carrier frequency [SkCK05].

Another experiment under static conditions was conducted by Marques *et al.* [MPKZ01]. They determined the coherence time at 1.7 GHz, bandwidth 30 MHz, in Duisburg, with BS positions above rooftop and the MT stationary in the street, but with path lengths ranging from 20 to 320 m. Coherence was taken to be lost for correlation coefficients under 0.9, between frequency spectra of snapshots averaged over all available measurement-point pairs with the same time offset. The authors found median coherence period lengths between 4.3 and 6 s, depending on the location.

Characterisation of special outdoor environments

Richly scattering environments in an industrial setting were investigated by Kemp and Bryant, with Bluetooth applications in the 2.4 GHz Industrial, Scientific and Medical (ISM) band in mind [KeBr05]. Path lengths were therefore relatively short, about 20 m on a petrochemical plant, 80 m on a car park lined with multistorey buildings, and up to 250 m on an electricity distribution transformer station. Delay spreads and BER were determined from about 120,000 snapshots, the majority of them taken on the petrochemical plant. As the equipment had a higher transmit power than usual Bluetooth appliances, the impulse responses could be clipped at 30 dB instantaneous dynamic range during processing. The average delay spread on the petrochemical plant was 38 ns, that on the transformer station 88 ns, and on the car park 76 ns. The variance of the results on the car park was two orders of magnitude higher than of those on the other environments. The authors attribute this to the

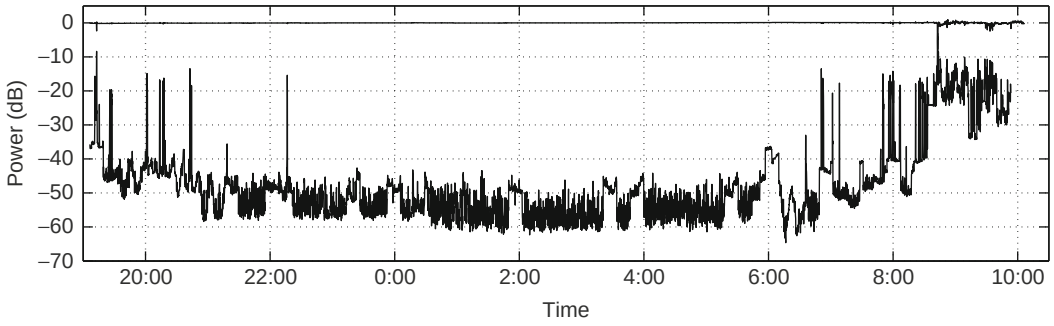
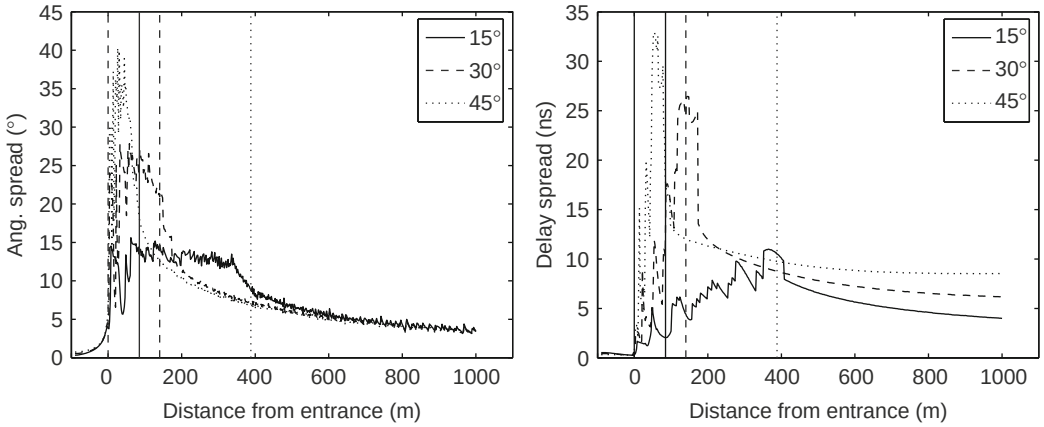


Figure 4.7-44 Total wideband power (upper curve) and time-varying power (lower curve) during night time (after [MeBH04]).



(a) Angular spread. (b) Delay spread.

Figure 4.7-45 Angular and delay spread in excitation zone behind tunnel entrance for three different Tx elevations: 15 °, 30°, and 45° (solid, dashed, and dotted). Tunnel entrance at 0 m; vertical lines indicate end of respective excitation zones (after [MoRJL03]).

presence of two distinctly different reflective processes, those of the cars in the direct vicinity, measured in metres, and those of the buildings at about 100 m distance, measured in hundred metres [KeBr05].

In comparison, indoor environments at 60 GHz would be easier to model, as reported by Geng *et al.* [GKZV05]. From the measured directional patterns at the campus of Helsinki University of Technology, they concluded that in LoS cases, the channel is well described by the direct path and first-order reflections. In NLoS cases, diffraction is the main propagation mechanism.

A special case of non-stationary propagation conditions is encountered in tunnels, just behind the entrance. Molina-García-Pardo *et al.* investigated this transitional regime, called 'excitation zone' [MaLD94], using UTD RT to examine path loss, delay spread, and angular spread over the first hundreds of metres in a rectangular straight tunnel of cross-section $8.5 \times 5.2\ \mathrm{m}^2$ ($w \times h$) [MoRJL03]. The transmitter, at 2.1 GHz, was placed at the tunnel's centreline at 100 m distance from the entrance at three different heights, corresponding to elevations of 15°, 30°, and 45°. In the excitation zone, delay spread, angular spread and path loss change drastically. The angular spread reaches it maximum transitional value at the beginning of this excitation zone, the delay spread at the end, Figure 4.7-45. The length of the excitation zone depends on the elevation angle of the transmitter as do the maximum transitional values of the spreads: the smaller the elevation, the longer the excitation zone, and the smaller the maximum transitional spreads.

4.7.5.8 Fading prediction

Introduction

In mobile radio system engineering it would be an advantage if channel fading behaviour could be forecast, using measurements of past and present behaviour. Predicting fading behaviour has promising applications in controlling relatively simple combining schemes in cheap receivers where knowing the occurrences of fades beforehand could be used to advantage [AJJF99]. Another envisaged use is that of transmit diversity in situations with insufficient knowledge of the channel or with too low update rates of the channel state information [DuHH00]. Predicting the channel state could then alleviate the need for frequent channel state updates.

This idea has been taken up by several authors. The early publications mainly treat the narrowband case [EyDHH98], [HwWi98], [AJJF99] but later ones also address wideband aspects [SeKa02b]. A number of publications showed a limited achievable prediction range, normally fractions of a wavelength to about one wavelength under favourable conditions ([AJJF99], [Teal01], [Ekma02], [SeKa03]).

Principle of fading prediction

As an example of the principle of fading prediction, we will treat here the narrowband case. Extensions to the broadband case, to be discussed later on, were only recently made by Semmelrodt and Kattenbach [SeKa02b]. A common approach of fading prediction is that of frequency analysis on the complex fading signal, for instance, by the use of high-resolution spectral estimation schemes like ESPRIT or root-MUSIC. The constituting sinusoidal signals are easily extrapolated and the predicted signal is formed by the same linear combination of the now extrapolated sinusoids as in the original signal, Figure 4.7-46. The attractiveness of this approach lies in the fact that numerous frequency analysis tools are available, see for instance [StMo97], and in the physical interpretation of the results. The sinusoidal components are readily identified as Doppler components [AJJF99],

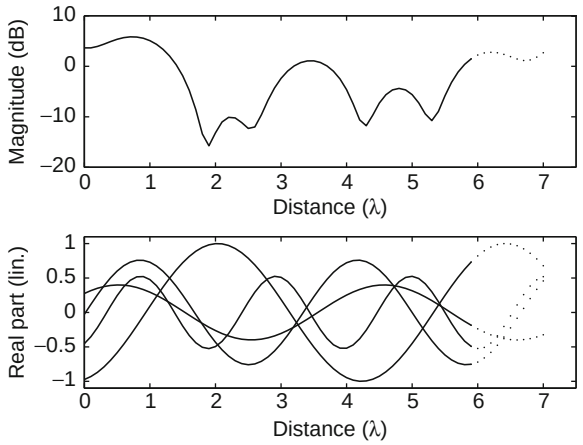


Figure 4.7-46 Principle of fading prediction based on spectral estimators. The measured complex fading signal (top, solid, magnitude shown) is decomposed in complex sinusoidal components (bottom, solid, real part shown). These are extrapolated beyond the measurement interval (bottom, dotted) and combined into the fading signal prediction (top, dotted).

stemming from important reflections in the environment with different angles of arrivals. In the case of narrowband prediction, the individual delays of the components are only visible as additional phase shifts. In practice, this approach does not model the numerous small components attributed to diffuse scattering, the curvature of the wavefronts of the Doppler components, and non-stationary channel states.

Narrowband fading prediction

Semmelrodt and Kattenbach made an extensive survey of spectral estimators for prediction algorithms [SeKa03]. They compared spectral estimators based either on spectral search (MUSIC and Modified Covariance), parameter estimation (EM and RELAXation spectral estimator (RELAX)), and subspace parameter estimation (ESPRIT, Unitary ESPRIT, root-MUSIC) or on adaptive filter algorithms, derived from autoregressive modelling. The adaptive filter algorithms comprised Modified Covariance, Least Mean Squares algorithm (LMS), Normalised Least Mean Squares algorithm (NLMS), Affine Projection (AfP), Recursive Least Squares algorithm (RLS), and Q-R decomposition-based Recursive Least Squares algorithm (QR-RLS). The comparison was made on synthetic data and measured data. Synthetic data was generated according to either the well-known Jakes model, using nine scatterers giving a small set of deterministic well-separated Doppler shifts, or the stochastic Dersch-model developed for the Research and Technology Development in Advanced Communication Technologies (RACE) Code Division Testbed RACE Project (CoDiT) project that only specifies temporal correlation properties, not Doppler profiles [BrDe91]. The measurements were recorded in a hall of $15 \times 15 \times 7\ \mathrm{m}^3$ ($l \times w \times h$) at 5.2 GHz over a distance of 1.84 m. The data was filtered in the Doppler domain in order to suppress out-of-band noise.

Figure 4.7-47 shows that the best performance is achieved with the Modified Covariance method for the stochastic synthetic channels and the measured channels, not with Jakes channels (not shown). The prediction length shown is defined as the length for which the relative mean-square error between the exact and predicted fading signal reaches −20 dB [SeKa03] and is normalised with respect to the wavelength. The QR-RLS method showed a performance close to that of the Modified Covariance method. For Jakes channels, subspace-based parameter estimation schemes like Unitary ESPRIT and root-MUSIC are preferred. They also perform well with stochastic (outdoor) channels with relatively few and separated scatterers but showed mediocre performance with the indoor measurements. Schemes that work on a fixed delay and Doppler grid, like EM-based methods, do in general not have enough resolution to produce any useful prediction. Therefore, Semmelrodt and Kattenbach suggest that the adaptive filter approach is preferred, giving the best results in realistic scenarios and being of (significantly) less computational complexity than the spectral estimators [SeKa03]. Also, extension to wideband prediction schemes is relatively straightforward.

Error sources in narrowband fading prediction

Kotterman investigated the accuracy of narrowband fading prediction for synthetic channels, consisting of deterministic Doppler components of random shifts, with prediction algorithms built around subspace-based parameter estimator schemes like ESPRIT, root-MUSIC, and some derivatives [Kott04]. His conclusion was that the strongest influence on the achievable extrapolation length for these types of channel came from the relation between the length of the observation interval over which the data is known and the actual number of scatterers. Channels with relatively high densities of scatterers or diffusely scattered components pose severe problems for the subspace-based spectral estimation schemes, see also Besson and Stoica [BeSt00], demanding impractically long data interval lengths over which stationarity of the channel cannot be guaranteed. As a rule of thumb, for a prediction length of one wavelength, the length of the data interval in wavelengths should equal at least half the number of scatterers in the channel, for noise-free cases. The fundamental reason for this is that the well-known reciprocal relation between resolution in time and that in frequency limits the accuracy of the Doppler estimation on short intervals. The maximum achievable prediction length depends mainly on this accuracy. Besides, one of the consequences is that the best prediction schemes for short data intervals render non-physical channel states [Kott04]. As additive noise reduces the achievable

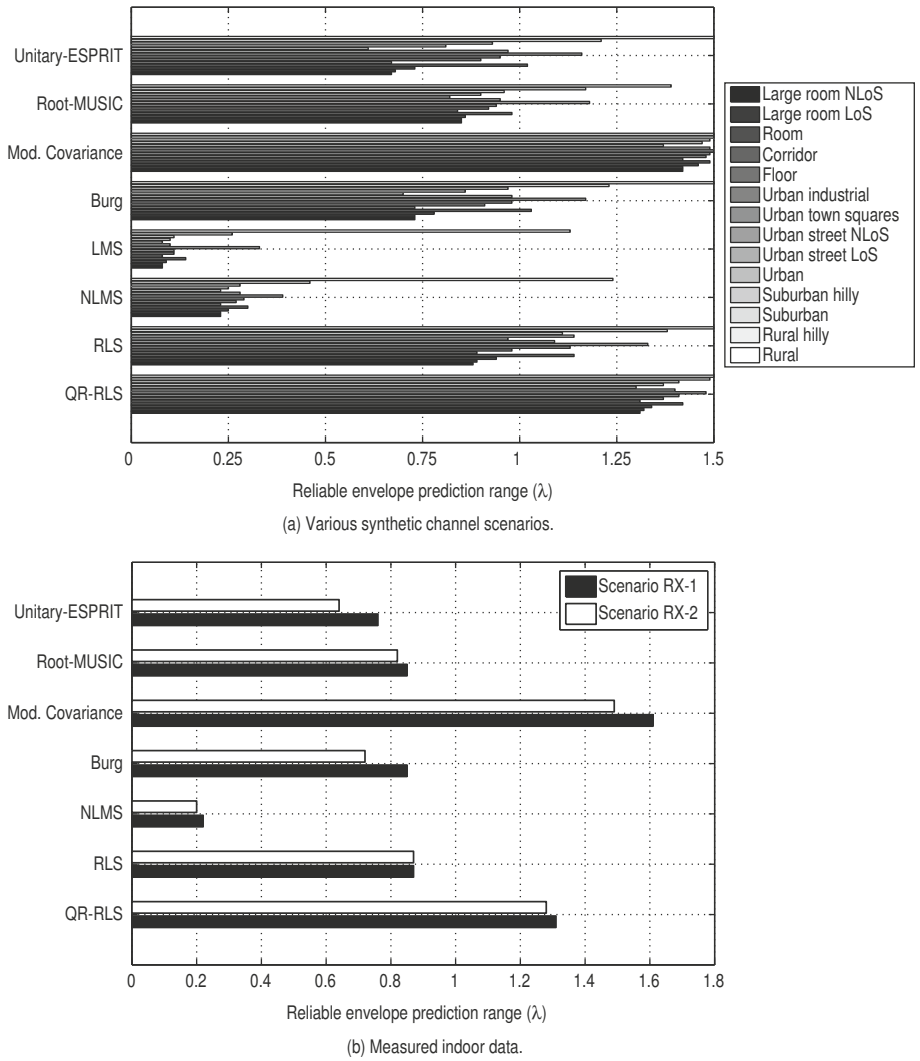


Figure 4.7-47 Extrapolation lengths for various synthetic channel scenarios and for measured indoor data, for prediction schemes based on Unitary ESPRIT, root-MUSIC and several AR-based schemes (see text) (after [SeKa03]).

prediction length too, practical prediction lengths with prediction schemes based on subspace spectral estimators will be limited to less than a wavelength, a conclusion earlier reached by Teal [Teal01].

In this respect, Jakes models with nine (angularly) well-separated deterministic Doppler components are ideally suited for subspace-based spectral estimators, as already noted by Semmelrodt and Kattenbach [SeKa03], and using these models for evaluation of fading prediction performance [SeKa03], [DuHH00], [Ekma02] is likely to give far too optimistic results as compared to realistic channels with dense scattering [Kott04].

Wideband fading prediction

In case of non-flat fading channels, narrowband extrapolators cannot be used over the entire bandwidth and the wideband character of the channel must be taken into account. Semmelrodt originally pioneered real broadband processing by taking the frequency dependence of the Doppler shifts and the time dependence of the delays into account [Semm04b]. In his two-point analysis, these dependences were estimated by applying a 2D ESPRIT routine twice, at different measurement times. Later on even a 4D SAGE was used on a slightly different formulation, estimating delay, Doppler, and the time gradients of both [Semm04b]. These approaches were abandoned with the success of the much simpler AR-based prediction on narrowband channels described above. Semmelrodt investigated three extensions to the flat fading approach [Semm04b]:

1. using 1D estimation/prediction algorithms in parallel, one per frequency channel;
2. combining information from multiple frequency channels to derive a single extrapolation scheme for the full bandwidth;

3. using 2D Doppler-delay estimation, combining all frequency channels.

Approach 1 is inherently able to treat frequency-dependent Doppler shifts but the frequency resolution per single-frequency prediction is not better than in the narrowband case. The other two approaches sacrifice frequency-dependent information for better resolution. With approach 2, more realisations of the same process are available for the estimation, assuming that the signal bandwidth is much larger than the coherence bandwidth of the channel. Approach 3 makes potentially the best use of the separation between paths in the delay-Doppler domain, but only under stationary channel conditions.

On synthetic channels with five deterministic Doppler components, the Unitary-ESPRIT based method showed better performance than the AR-based Modified Covariance, but in relation to the three wideband approaches the relative results of both methods were comparable: the parallel 1D approach showed the worst performance, similar to the narrowband approximation. The 2D approach showed the best results with approach 2 in between, closer to the 2D than to the 1D results.

Figure 4.7-48 shows the prediction results for the measured indoor data presented in the section narrowband fading prediction, with a bandwidth of 120 MHz. Again, a mean-square error criterion is used, set at −20 dB. Prediction lengths with root-MUSIC were consistently better than half a wavelength (no 2D algorithm was available for root-MUSIC). Unitary ESPRIT performed worst, only with approach 2 ('Multi-Channel 1D') a prediction length of about one wavelength was achieved whereas the narrowband approximation and the 2D approach failed. In these cases, Semmelrodt expects better results when the model order of the ESPRIT estimation can be made adaptive. Best performance was achieved with the AR-based method; in all three approaches a prediction length of about one and a half wavelength was achieved. The estimation errors were constant over the bandwidth and were comparable for the three approaches [Semm04b].

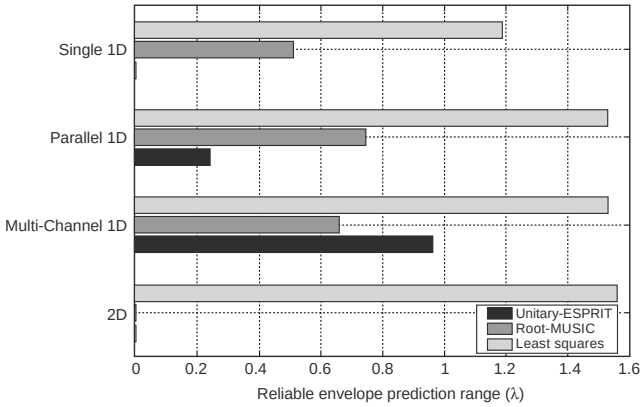


Figure 4.7-48 Comparison of prediction performance of three prediction schemes on measured indoor data for three wideband approaches: 'Parallel 1D', 'Multi-channel 1D', and '2D', see text (after [Semm04b]).

References

[AJJF99] J. B. Andersen, J. Jensen, S. Jensen, and F. Frederiksen. Prediction of future fading based on past measurements. In *Proc. VTC1999 Fall – IEEE 51st Vehicular Technology Conf.*, Amsterdam, The Netherlands, Sep. 1999.

[AlRT04] V. Algeier, A. Richter, and R. Thomä. A gradient based algorithm for path parameter tracking in channel sounding. *COST 273*, TD(04)124, Gothenburg, Sweden, June 2004.

[Ande02] J. A. Andersen. Power distributions revisited. *COST 273*, TD(02)004, Guildford, UK, Jan. 2002.

[AnMd01] M. R. Andrew, P. P. Mitra, and R. deCarvalho. Tripling the capacity of wireless communications using electromagnetic polarization. *McMillan Magazines*, 409: 316–318, Jan. 2001.

[BCFF02] M. Barbiroli, C. Carciofi, G. Falciasecca, M. Frullone, P. Grazioso, and A. Varini. A new statistical approach for urban environment propagation modelling. *IEEE Trans. Veh. Technol.*, 51(5): 1234–1241, Sep. 2002. TD-02-044.

[Berg95] J. E. Berg. A recursive method for street microcell path loss calculations. In *Proc. of 6th IEEE Intern. Symp. on Personal, Indoor and Mobile Radio Communications*, Toronto, Canada, Sep. 1995.

[Bert02] H. L. Bertoni. Status of ray tracing codes with application to Monte Carlo simulation of channel parameters. *COST 273*, WP(02)016, Espoo, Finland, May 2002.

[BeSt00] O. Besson and P. Stoica. Decoupled estimation of DoA and angular spread for a spatially distributed source. *IEEE Trans. Signal Processing*, 48(7): 1872–1882, July 2000.

[BHHT04a] H. Budiarto, K. Horihata, K. Haneda, and J. Takada. Experimental study of non-specular wave scattering from building surface roughness for the mobile propagation modeling. *IEICE Trans. Commun.*, E87-B(7):958–966, Apr. 2004. [Also available as TD(03)200.]

[BHHT04b] H. Budiarto, K. Horihata, K. Haneda, and J. Takada. Experimental study of non-specular wave scattering from building surface roughness for the mobile propagation modeling. *IEICE Trans. Commun.*,

E87-B(4):958–966, Apr. 2004. [Also available as TD(03)200.]

[BrCu01] R. Bradley and P.J. Cullen. The effect of transverse surface gradients on propagation over undulating terrain, modelled with a PEC surface integral equation. *COST 273*, TD(01)006, Brussels, Belgium, May 2001.

[BrDe91] W. R. Braun and U. Dersch. A physical mobile radio channel model. *IEEE Trans. Veh. Technol.*, 40(2):472–482, May 1991.

[Bult02a] R. J. C. Bultitude. A comparison of multipath-dispersion-related microcellular mobile radio channel characteristics at 1.9 GHz and 5.8 GHz. In *Proc. Symp. Antenna Techn. Appl. Electromagnetics (ANTEM 2002)*, July 2002. [Also available as TD-03–015.]

[Bult02b] R. J. C. Bultitude. Estimating frequency correlation functions from propagation measurements on fading radio channels: a critical review. *IEEE J. Select. Areas Commun.*, 20(6): 1133–1143, Aug. 2002.

[Bult03] R. J. C. Bultitude. Segmentation of measured data and modelling of the nonstationary characteristics of narrowband radio channels in urban microcells. *COST 273*, TD(03)135, Paris, France, May 2003.

[Bult04] R. J. C. Bultitude. Considerations concerning statistical stationarity in the analysis of measured data and radio channel modelling. *COST 273*, Duisburg, Germany, Sep. 2004. [Tutorial at COST 273.]

[CaCo01] F. Cardoso and L. M. Correia. An analytical approach to fading depth dependence on bandwidth for mobile communication systems. In *Proc. WPMC2001 – Wireless Pers. Multimedia Commun.*, Aalborg, Denmark, Sep. 2001. [Also available as TD(01)025.]

[CaCo02] F. Cardoso and L. M. Correia. A time-domain technique for fading depth characterisation in wideband mobile communications systems. In *Proc. of URSI General Assembly*, Maastricht, The Netherlands, Aug. 2002. [Also available as TD(02)038.]

[CaCo03] F. Cardoso and L. M. Correia. A comparison between different approaches for fading evaluation in wideband mobile communications. In *Proc. VTC 2003 Spring – IEEE 57th Vehicular Technology Conf.*, Jejou, South Korea, Apr. 2003. [Also available as TD(03)046.]

[CaCo04] F. Cardoso and L. M. Correia. Short-term fading depth dependence on antenna characteristics in wideband mobile communications. In *Proc. PIMRC 2004 – IEEE 15th Int. Symp. on Pers., Indoor and Mobile Radio Commun.*, Barcelona, Spain, Sep. 2004. [Also available as TD(04)089.]

[Card04] F. D. Cardoso. *Short-term Fading Characterisation in Wideband Mobile Communication Systems*. PhD thesis, Instituto Superior Técnico, Technical University of Lisbon, Lisbon, Portugal, 2004.

[Carl86] A. B. Carlson. *Communication Systems – An Introduction to Signals and Noise in Electrical Communication*. McGraw-Hill, Singapore, 1986.

[CBYF05] N. Czink, E. Bonek, X.-F. Yin, and B. Fleury. Cluster angular spreads in a MIMO indoor propagation environment. In *14th IST Mobile & Wireless Communications Summit*, Dresden, Germany, June 2005.

[ChBe01] G. Liang, C. Cheon, and H. L. Bertoni. Simulating radio channel statistics for different building environments. *IEEE J. Select. Areas Commun.*, 19(11): 2191–2200, 2001.

[CHÖB04] N. Czink, M. Herdin, H. Özcelik, andE. Bonek. Cluster characteristics in a MIMO indoor propagation environment. *COST 273*, TD(04)167, Duisburg, Germany, Sep. 2004.

[ChSA04] H. Y. E. Chua, K. Sakaguchi, and K. Araki. Experimental and analytical investigation of MIMO channel capacity in an indoor LOS environment. *COST 273*, TD-04-023, Jan. 2004.

[ChWa92] W. C. Chew and R.L. Wagner. A modified form of Liao's absorbing boundary condition. In *IEEE AP-SInt., Symp., Dig.*, pages 536–539, 1992.

[CoLA03] Y. Corre, Y. Lostanlen, and S. Aubin. Multi-environment radio predictions involving an in-building WLAN network and outdoor UMTS base stations. *COST 273*, TD(03)156, Prague, Czech Republic, Sep. 2003.

[Corr01] L. M. Correia. *Wireless Flexible Personalised Communications, COST 259: European Co-operation in Mobile Radio Research*. John Wiley & Sons, Chichester,UK, 2001.

[Cost03] F. Costen. Analysis and improvement of Liao ABC for FDTD. In *AP-S Int. Symp. USNC/CNC/URSIRadio Science Meeting*, pages 23–25, Columbus, OH, USA, June 2003.

[Cost04] F. Costen. A proposal of 3D frequency dependent ADI-FDTD with conductive loss. *In Int. Conf. Comput. Electromagn.*, pages 169–170, 2004.

[CoTh04a] F. Costen and A. Thiry. Alternative formulation of three dimensional frequency dependent ADI-FDTD method. *IEICE Electron.* Express, 1: 528–533, 2004.

[CoTh04b] F. Costen and A. Thiry. Temporal discretization for UWB systems in three dimensional alternating-direction implicit finite difference time domain method. *IEICE Electron. Express*, 1:477–483, 2004.

[CoWS04] M. Coinchon, J. F. Wagen, and A.P. Salovaara. Joint channel estimation with array antennas in OFDM based mobile radio systems. *COST 273*, TD(01)041, Bologna, Italy, Oct. 2004.

[Cox73] D. C. Cox. 910 MHz urban mobile radio propagation: multipath characteristics in New York City. *IEEE Trans. Commun.*, 21(11): 1188–1194, Nov. 1973.

[CTLM02] C.-C. Chong, C.-M. Tan, D. I. Laurenson, S. McLaughlin, M.A. Beach, and A.R. Nix. Indoor wideband directional channel measurement and

modelling. In *1st COST273 Workshop 'Opportunities of the multidimensional propagation channel'*, Helsinki, Finland, May 2002.

[CzYi05] N. Czink and X.-F. Yin. Cluster angular spread estimation for MIMO indoor propagation environments. *COST 273*, TD(05)041, Bologna, Italy, Jan. 2005.

P. Daniele, V. Degli-Esposti, G. Falciasecca, and G. Riva. Field prediction tools for wireless communications in outdoor and indoor environments. In *IEEE MTT-S European Topical Congress 'Technologies for Wireless Applications'*.

[DDGW03] D. Didascalou, M. Döttling, N. Geng, and W. Wiesbeck. An approach to include stochastic rough surface scattering into deterministic ray-optical wave propagation modeling. *IEEE Trans. Antennas Propagat.*, 51(7): 27–37, July 2003. [Also available as TD(01)007.]

[DeBe99] V. Degli-Esposti and H.L. Bertoni. Evaluation of the role of diffuse scattering in urban microcellular propagation. In *Proc. VTC 1999 Fall – IEEE 48th Vehicular Technology Conf.*, Amsterdam, The Netherlands, Sep. 1999.

[DEESGK02a] V. Degli-Esposti, H. El-Sallabi, D. Guiducci, K. Kalliola, P. Azzi, L. Vuokko, J. Kivinen, and P. Vainikainen. Analysis and simulation of the diffuse scattering phenomenon in urban environment. In *Proc. of the XXVII URSI General Assembly*, Maastricht, The Netherlands, Aug. 2002. [Also available as TD(02)036.]

[DEESGK02b] V. Degli-Esposti, H. El-Sallabi, D. Guiducci, K. Kalliola, P. Azzi, L. Vuokko, J. Kivinen, and P. Vainikainen. Analysis and simulation of the diffuse scattering phenomenon in urban environment. In *Proc. of the XXVIIth URSI General Assembly*, Maastricht, The Netherlands, Aug. 2002. [Also available as TD-02-036.]

[DeFA04] V. Degli-Esposti, F. Fuschini, and M. Amorini. Database simplification for field prediction in urban environment. In *IEEE AP-S/ URSI Int. Symposium*, Monterrey, CA, USA, June 2004. [Also available as TD(04)041.]

[DeFG03] V. Degli-Esposti, F. Fuschini, and D. Guiducci. A study on roof-to-street propagation. In *Proc. ICEAA 2003-IEEE Int. Conf. on Electromagnetics in Advanced Applications*, Turin, Italy, Sep. 2003. [Also available as TD(03)143.]

[DEFVG05] V. Degli-Esposti, F. Fuschini, E. Vitucci, and D. Graziani. Measurement and modelling of diffuse scattering from building walls. *COST 273*, TD(05)065, Bologna, Italy, Sep. 2005.

[DEGDA04] V. Degli-Esposti, D. Guiducci, A. De Marsi, P. Azzi, and F. Fuschini. An advanced field prediction model including diffuse scattering. *IEEE Trans. Antennas Propagat.*, 14(7): 1717–1728, July 2004.

[Degl01] V. Degli-Esposti. A diffuse scattering model for urban propagation prediction. *IEEE Trans. Antennas Propagat.*, 49(7): 1111–1113, July 2001.

[Deyg66] J. Deygout. Multiple knife edge diffraction of microwaves. *IEEE Trans. Antennas Propagat.*, 14(4): 480–489, July 1966.

[DMSW03] D. Didascalou, J. Maurer, T. Schäfer, and W. Wiesbeck. An approach to include stochastic rough surface scattering into deterministic ray-optical wave propagation modelling. *IEEE Trans. Antennas Propagat.*, 51(7): 27–37, July 2003. [Also available as TD(01)007.]

[DuHH00] A. Duel-Hallen, S. Hu, and H. Hallen. Long-range prediction of fading signals. *IEEE Signal Processing Mag.*, 17(3): 62–75, May 2000.

[EFLR97] V. Erceg, S. J. Fortune, J. Ling, A. J. Rustako, and R. A. Valenzuela. Comparisons of a computer-based propagation prediction tool with experimental data collected in urban microcellular environments. *IEEE J. Select. Areas Commun.*, 15(4): 677–684, May 1997.

[EFVG05] V. Degli-Esposti, F. Fuschini, E. Vitucci, and D. Grazian. Measurement and modelling of scattering from building walls. *COST 273*, TD(05)065, Bologna, Italy, Jan. 2005.

[Ekma02] T. Ekman. *Prediction of mobile radio channels, modelling and design*. PhD thesis, Uppsala University, Uppsala, Sweden, 2002. ISBN 91-506-1625-0.

[ElVa03] H. El-Sallabi and P. Vainikainen. A new heuristic UTD diffraction coefficient for prediction of radio wave propagation. *COST 273*, TD(03)027, Guildford, UK, Jan. 2003.

[ETSI97] ETSI. *Universal mobile telecommunications system (UMTS), selection procedures for the choice of radio transmission technologies of the UMTS*. UMTS 30.03, version 3.1.0, Sophia Antipolis, France, Nov. 1997.

[EyDHH98] T. Eyceoz, A. Duel-Hallen, and H. Hallen. Deterministic channel modelling and long range prediction of fast fading mobile channels. *IEEE Commun. Lett.*, 2(9): 254–256, Sep. 1998.

[FBKE02] S. E. Foo, M. A. Beach, P. Karlsson, P. Eneroth, B. Lindmark, and J. Johansson. Spatio-temporal investigation of UTRA FDD channels. In *3rd Intern. Conference on 3G Mobile Comm. Technol.*, May 2002. [Also available as TD-02–027.]

[FGLM04] R. Fraile, J. Gozálvez, O. Lázaro, J. Monserrat, and N. Cardona. Effect of a two dimensional shadowing model on system level performance evaluation. In *Proc. WPMC 2004-Wireless Pers. Multimedia Commun.*, Abano Terme, Italy, Sep. 2004. [Also available as TD(04)190.]

[FlSJ02] B. Fleury, A. Stucki, and P. Jourdan. High resolution bidirection estimation based on the SAGE algorithm: Experience gathered from field experiments. *COST 273*, TD(02)070, Espoo, Finland, May 2002.

[FMDW03] T. Fügen, J. Maurer, A. Dallinger, and W. Wiesbeck. Radio channel characterization with ray-tracing for urban environments at 2 GHz. *COST 273*, TD(03) 130, Paris, France, May 2003.

[FrLC03] R. Fraile, O. Lazaro, and N. Cardona. Two dimensional shadowing model. *COST 273*, TD(03)171, Prague, Czech Republic, Sep. 2003.

[FuMa03] T. Fuegen and J. Maurer. Radio channel characterization with ray-tracing for urban evironment at 2 GHz. *COST 273*, TD(03)130, Paris, France, May 2003.

[Fusc04] F. Fuschini. A study on urban gap fillers for DVB-H system in urban environment. *COST 273*, TD(04)201, Duisburg, Germany, Sep. 2004.

[FYSJ03] B. Fleury, X. Yin, A. Stucki, and P. Jourdan. High resolution channel parameter estimation for communication systems equipped with antenna arrays. *COST 273*, TD(03)132, Paris, France, May 2003.

[GaHS03] G. Del Galdo, M. Haardt, and C. Schneider. Ilmprop: a flexible geometry-based simulation environment for multiuser MIMO communications. In *Advances in Radio Science – Kleinheubacher Berichte*, page 10, Oct. 2003. [Also available asTD(02)188.]

[GCBV02] P. Grazioso, C. Carciofi, M. Barbiroli, and A. Varini. A statistical propagation model for urban environment. *COST 273*, TD(02)044, Espoo, Finland, May 2002.

[Geng05] S. Geng. Propagation characterization of 60 GHz indoor radio channels. *COST 273*, TD(05)015, Bologna, Italy, Jan. 2005.

[GhTI04] M. Ghoraishi, J. Takada, and T. Imai. Investigating dominant scatterers in urban mobile propagation channel. In *Proc. of ISCIT 2004 – IEEE Int. Symp. on Communications and Information Technologies*, Sapporo, Japan, Oct. 2004. [Also available as TD(04)154.]

[GKZV05] S. Geng, J. Kivinen, X. Zhao, and P. Vainikainen. Propagation characterisation of 60GHz indoor radio channels. *COST 273*, TD(05)015, Bologna, Italy, Jan. 2005.

[GLBM02] L. Goossens, E. Van Lil, B. Boesmans, and D. Magnus. Wind turbine sites in industrial areas: specific problems and solutions. In *Proc. of Powergen Europe*,Milano, Italy, June 2002. [Also available as TD(02)124.]

[GRBM03] S. G. Garcia, R. G. Rubio, A. R. Bretones, and R. G. Martin. Extension of the ADI-FDTD method to Debye media. *IEEE Trans. Antennas Propagat.*, 51-11: 3183–3186, 2003.

[Hans02] J. Hansen. Towards a geometrically motivated, analytical indoor channel model. *COST 273*, TD(02)063, Espoo, Finland, May 2002.

[Hans03] J. Hansen. Analytical calculation of the power delay profile and delay spread with experimental verification. *IEEE Commun. Lett.*, 7(6): 257–259, 2003. [Also available as TD-02-010.]

[Hata80] M. Hata. Empirical formula for propagation-loss in land mobile radio service. *IEEE Trans. Veh. Technol.*, 29(3): 317–325, Aug. 1980.

[HaTa03a] K. Haneda and J. Takada. An application of the SAGE algorithm for UWB propagation channel estimation. In *Proc. of UWBST 2003 – 2nd IEEE Conf. on Ultra Wideband Systems and Technologies*, Reston, VA, USA, Nov. 2003. [Also available as TD(03)182.]

[HaTa03b] K. Haneda and J. Takada. High resolution estimation of the NLOS indoor MIMO channel with a network analyser based system. In *Proc. PIMRC 2003 – IEEE 14th Int. Symp. on Pers., Indoor and Mobile Radio Commun.*, Beijing, China, Sep. 2003. [Also available as TD(03)119.]

[HaTK05] K. Haneda, J. Takada, and T. Kobayashi. On the cluster properties in UWB spatio-temporal residential measurement. *COST 273*, TD(05)066, Bologna, Italy, Jan. 2005.

[HeKu05] A. Hecker and T. Kuerner. Analysis of propagation models for UMTS ultra high sites in urban areas. *COST 273*, TD(05)033, Bologna, Italy, Sep. 2005.

[Herd04] M. Herdin. *Non-Stationary Indoor MIMO Radio Channels*. PhD thesis, Vienna University of Technology, Vienna, Austria, 2004. [Also available as TD(04) 174.]

[HHBW01] R. Hoppe, H., Buddendick, G. Wölfle, and F. M. Landstorfer. Dynamic simulator for studying WCDMA radio network performance. In *Proc. VTC 2001 Spring – IEEE 53rd Vehicular Technology Conf.*, Rhodos, Greece, May 2001. [Also available as TD(02) 148.]

[HMDA02] C. L. Holloway, P. M. McKenna, R. A. Dalke, and C. L. Perala Devor Jr. Time-domain modeling, characterization, and measurements of anechoic and semi-anechoic electromagnetic test chambers. I*EEE Trans. Electromagn. Compat.*, 44: 102–118, 2002.

[HMVT94] H. Hashemi, H. McGuire, M. Vlasschaert, T. Tholl, and D. Tholl. Measurements and modelling of temporal variations of the indoor radio propagation channel. *IEEE Trans. Veh. Technol.*, 43(3): 733–737, Aug. 1994.

[HoWW03] R. Hoppe, G. Woelfle, and P. Wertz. Advanced ray-optical wave propagation modelling for urban and indoor scenarios. *European Transactions on Telecommunications (ETT)*, 14(1): 61–69, Jan. 2003. [Also available as TD(02)051.]

[HuKL02] K. Hugl, K. Kalliola, and J. Laurila. Spatial channel characteristics for macro and microcellular BS installations. In *1st COST 273 Workshop 'Opportunities of the multidimensional propagation channel'*, Helsinki, Finland, May 2002. COST 273.

[HwWi98] J. Hwang and J. Winters. Sinusoidal modelling and prediction of fast fading processes. In *Proc. Globecom 1998-IEEE Global Telecommunications Conf.*, Sydney, Australia, Nov. 1998.

[JäHH04] T. Jämsä, V. Hovinen, and L. Hentilä. Comparison of wideband and ultrawideband channel measurements. *COST 273*, TD(04)080, Gothenburg, Sweden, June 2004.

[JePK05] J. Jemai, R. Piesiewicz, and T. Kuerner. Calibration of an indoor ray-tracing propagation model at 2.4 GHz by measurements of the IEEE 802.1 1b pre-amble. *In Proc. VTC2005 Spring – IEEE 61st Vehicular Technology Conf.*, Stockholm, Sweden, May 2005. [Also available as TD(04)149.]

[JKVW05] J. Jemai, M. Kürner, A. Varone, and J.-F. Wagen. Determination of the permittivity of building materials through WLAN meaasurements at 2.4 GHz. *COST 273*, TD(05)032, Bologna, Italy, Jan. 2005.

[KaFr01] R. Kattenbach and H. Früchting. Investigation of the impacts of moving scatterers by wideband measurements of time-variant indoor radio channels. *Frequenz*, 55(7–8): 197–203, July 2001. In German. [Also available as TD(01)033.]

[KaKC02] P. D. Karamalis, A. G. Kanatas, and P. Constantinou. A genetic algorithm applied for optimization of antenna arrays used in mobile radio channel characterization devices. *COST 273*, TD(02)091, May 2002.

[Katt97] R. Kattenbach. *Characterisation of time-variant indoor radio channels by means of their system and correlation functions*. PhD thesis, Universität Kassel, Kassel, Germany, 1997. In German, ISBN 3-8265-2872-7.

[Katt02] R. Kattenbach. Transfer function modeling and its application to ultra-wideband channels. *COST 273*, TD(02)136, Lisbon, Portugal, Sep. 2002.

[KaVV05] A. Kainulainen, L. Vuokko, and P. Vainikainen. Polarization behaviour in different urban radio environments. *COST 273*, TD(05)018, Bologna, Italy, Jan. 2005.

[KaWe00] R. Kattenbach and D. Weitzel. Wideband channel sounder for time-variant indoor radio channels. In *Proceedings AP2000 Millennium Conference on Antennas and Propagation*, Davos, Switzerland, Apr. 2000.

[KeBr05] A. H. Kemp and E. B. Bryant. Channel sounding of industrial sites in the 2.4 GHz ISM band. *Wireless Personal Communications*, 31(3–4): 235–248, Dec. 2004. [Also available as TD-02-111.]

[KeCu03] E. Kenny and P. J. Cullen. An analytical formula for obtaining the lookup table in the tabulated interaction method (TIM) for rough terrain. *COST 273*, TD(03) 139, Paris, France, May 2003.

[KEGJ03] T. Kuerner, A. Eisenblätter, H.-F. Geerdes, D. Junglas, T. Koch, and A. Martin. *Final report on automatic planning and optimisation, technical report IST- 2000-28088-MOMENTUM-D47-PUB*. Deliverable D4.7 of IST-MOMENTUM, European Commission, Brussels, Belgium, 2003. [Also available as TD(03)172.]

[KKVV04] V.-M. Kolmonen, J. Kivinen, L. Vuokko, and P. Vainikainen. 5.3 GHz MIMO radio channel sounder. *COST 273*, TD-04–141, June 2004.

[KKVV05] V.-M. Kolmonen, J. Kivinen, L. Vuokko, and P. Vainikainen. 5.3 GHz MIMO radio channel sounder. In *Proc. IMTC '05-IEEE Instrumentation and Measurement Technology Conf.*, submitted for publication, 2005. [Also available as TD(04)193.]

[Kott04] W. A. Th. Kotterman. *Characterisation of Mobile Radio Channels for Small Multiantenna Terminals*. PhD thesis, Aalborg Universitet, Aalborg, Denmark, 2004. ISBN 87-90834-68–2.

[KSPH01] J. Kivinen, P. Suvikunnas, D. Perez, C. Herrero, K. Kalliola, and P. Vainikainen. Characterization system for MIMO channels. In *Proc. WPMC 2001 – Wireless Pers. Multimedia Commun.*, Aalborg, Denmark, Oct. 2001. [Also available as TD(01)044.]

[KüMe02] T. Kürner and A. Meier. Prediction of outdoor and outdoor-to-indoor coverage in urban areas at 1.8 GHz. *IEEE J. Select. Areas Commun.*, 20(3): 496–506, Apr. 2002. [Also available as TD(01)013.]

[Kunn02] E. Kunnari. Statistical modeling of small-scale fading with temporal, spatial and spectral correlation and rice distributed envelope in the simulation of multi-carrier systems. *COST 273*, TD(02)019, Guildford, UK, Jan. 2002.

[Kyri03] P. Kyritsi. K factor estimation in a hallway using waveguide mode analysis. *COST 273*, TD(03)047, Barcelona, Spain, Jan. 2003.

[LAME05] J. Liu, B. Allen, W. Q. Malik, and D. J. Edwards. On the spatial correlation of MB-OFDM ultra wideband transmissions. *COST 273*, TD(05)015, Bologna, Italy, Jan. 2005.

[LaRT03] M. Landmann, R. Richter, and R. Thomä. Performance evaluation of real antenna arrays for high-resolution DoA estimation in channel sounding -part 1: channel parameter resolution limits. *COST 273*, TD(03)199, Sep. 2003.

[LaRT04a] M. Landmann, A. Richter, and R. S. Thomä. DoA resolution limits in MIMO channel sounding. In *2004 IEEE Intl. Symp. on Antennas and Propagation and USNC/URSI National Radio Science Meeting*, pages 20–26, Monterey, CA, USA, June 2004.

[LaRT04b] M. Landmann, A. Richter, and R. S. Thomä. Performance evaluation of antenna arrays for high-resolution DOA estimation in channel sounding. In *2004 Intl. Symp. on Antennas and Propagation*, pages 17–21, Sendai, JP, Aug. 2004.

[LBLC02] E. Van Lil, B. Van den Broeck, Y. Van Laer, and A. Van de Capelle. Validation of GO simulations with trunking measurements in Leuven. *COST 273*, TD(02)014, Guildford, UK, Jan. 2002.

[LCDC02] E. Van Lil, I. De Coster, Y. Demarsin, F. Casteels, and A. Van de Capelle. Finetuning the fresnel PO model. In *Proc. of the XXVIIth URSI General Assembly*, Maastricht, The Netherlands, Aug. 2002. [Also available as TD(02)084.]

[Lee85] W. Lee. Estimate of local average power of a mobile radio signal. *IEEE Trans. Veh. Technol.*, 34(1): 22–27, Feb. 1985.

[LiTC05] E. Van Lil, D. Trappeniers, and A. Van de Capelle. Simplified formulas for finding the extent of false echo zones in radar systems. *COST 273*, TD(05)039, Bologna, Italy, Jan. 2005.

[LKTH02] J. Laurila, K. Kalliola, M. Toeltsch, K. Hugl, P. Vainikainen, and E. Bonek. 3D characterization of mobile radio channels in urban environment. *IEEE Trans. Antennas Propagat.*, 50(2): 233–243, Feb. 2002.

[LoCo02a] Y. Lostanlen and Y. Corre. A 2. 5D model for predicting the in-building propagation penetration in indoor environments. *COST 273*, TD(02)052, Espoo, Finland, May 2002.

[LoCo02b] Y. Lostanlen and Y. Corre. Studies on indoor propagation at various frequencies for radio local networks. *COST 273*, WP(02)013, Espoo, Finland, May 2002.

[LoCo05] Y. Lostanlen and Y. Corre. Urban coverage simulations for broadcast (DVB-H, DVB-T) networks. *COST 273*, TD(05)048, Bologna, Italy, Sep. 2005.

[LoGC05] Y. Lostanlen, G. Gourgeon, and Y. Corre. An approach to model the ultra-wideband multipath indoor radio channel by ray-tracing methods. *COST 273*, TD(05)047, Bologna, Italy, Sep. 2005.

[LPEO02] P. H. Lehne, M. Pettersen, R. Eckhoff, and O. Trandem. A method for synchronising transmitter and receiver antenna switching when performing dual array measurements. *URSI General Assembly*, Aug. 2002. [Also available as TD(02)048.]

[LVTC04] E. Van Lil, J. Verhaevert, D. Trappeniers, and A. Van de Capelle. Theoretical investigations and broadband experimental verification of the time-domain SAGE DOA algorithm. In *20th Annual Review of Progress in Applied Computational Electromagnetics (ACES2004)*, Syracuse, NY, USA, Apr. 2004. [Also available asTD(04)142.]

[LZMJ05] D. Laselva, X. Zhao, J. Meinilä, T. Jämsä, J. Nuutinen, P. Kyösti, and L. Hentilä. Empirical large-scale characterization of the received power for rural, suburban and indoor environments at 2.45 and 5.25 GHz. *COST 273*, TD(05)043, Bologna, Italy, Jan. 2005.

[MaHe03] E. Martijn and M. Herben. Radio wave propagation into buildings at 1.8 GHz; empirical characterisation and its importance to UMTS radio planning. *COST 273*, TD(03)191, Prague, Czech Republic, Jan. 2003.

[MaLD94] Ph. Mariage, M. Lienard, and P. Degauque. Theoretical and experimental approach of the propagation of high frequency waves in road tunnels. *IEEE Trans. Antennas Propagat.*, 42(1): 75–81, Jan. 1994.

[MATB02] J. Medbo, H. Asplund, M. Törnqvist, D. Browne, and J.-E. Berg. MIMO channel measurements in an urban street microcell. In *Proc. of RVK'02 – Radio Vetenskap och Kommunikation*, Stockholm, Sweden, 2002. [Also available as TD(03)006.]

[MeBe02] J. Medbo and J. Berg. Simple and accurate path loss modeling at 5 GHz in complex indoor environments with corridors. In *Proc. URSI Commission F Open Symposium Radiowave Prop. Remote Sensing*, Garmisch-Partenkirchen, Germany, Feb. 2002. [Also available as TD(02)055.]

[MeBH04] J. Medbo, J.-E. Berg, and F. Harrysson. Temporal radio channel variations with stationary terminal. In *Proc. VTC 2004 Fall – IEEE 60th Vehicular Technology Conf.*, Los Angeles, CA, USA, Sep. 2004. [Also available as TD-0-4-183.]

[MGPCBRJL04] J.-M. Molina-Garcia-Pardo, J. Caldrán-Blaya, J.-V. Rodríguez, and Juan-Llácer. MIMO measurement system based on two network analyzers. *COST 273*, TD-04-106, June 2004.

[MiVV04] F. Mikas, L. Vuokko, and P. Vainikainen. Large scale behaviour of multipath fading channels in Urban macrocellular environments. *COST 273*, TD(04)101, Gothenburg, Sweden, June 2004.

[MoRJL03] J.-M. Molina-García-Pardo, J.-V. Rodríguez, and L. Juan-Llácer. Angular spread at 2.1 GHz while entering tunnels. *Microwave and Optical Technology Lett.*, 37(3): 196–198, May 2003. [Also available as TD-02-043.]

[MoRo04] S. Mota and A. Rocha. Experimental results from channel parameter estimation using the SAGE algorithm. *COST 273*, TD(04)138, Gothenburg, Sweden, June 2004.

[MPKZ01] G. Marques, J. Pamp, J. Kunisch, and E. Zollinger. *Wideband Directional Channel Model, Array Antennas and Measurement Campaign*. IST ASILUM project, Deliverable D4.3bis, Nov. 2001. [Also available as TD-02-095.]

[MuSV05] M. Mustonen, P. Suvikunnas, and P. Vainikainen. Reliability analysis of multidimensional propagation channel characterization. *COST 273*, TD(05)040, Bologna, Italy, Jan. 2005.

[NaCB04] K. Nasr, F. Costen, and S. K. Barton. A study of wall imperfections in indoor

channel modelling. *COST 273*, TD(04)017, Athens, Greece, Jan. 2004.

[Nami99] T. Namiki. A new FDTD algorithm based on alternating-direction implicit method. *IEEE Trans. Microwave Theory Tech.*, 47: 2003–2007, 1999.

[NdHH04] H. Ndoumbè Mbonjo Mbonjo, J. Hansen, and V. Hansen. Field theoretical investigations of the influence of mutual coupling effects on MIMO channels. In *Proc. Globecom 2004 – IEEE Global Telecommunications Conf.*, Dallas, TX, USA, Dec. 2004. [Also available as TD(03)145.]

[NePa94] A. Nehorai and E. Paldi. Vector-sensor array processing for electromagnetic source localization. *IEEE Trans. Signal Processing*, 42: 376–398, Feb. 1994.

[OOKF68] Y. Okumura, E. Ohmori, T. Kawano, and K. Fukuda. Field strength and its variability in VHF and UHF land mobile radio service. *Review of the Electrical Communication Laboratory*, 16(9–10): 825–873, Sep. 1968.

[Pars92] D. Parsons. *The Mobile Radio Propagation Channel*. Pentech Press, London, UK, 1992.

[PaTB04] A. Pal, C. Tan, and M. Beach. Comparision of MIMO channels from multipath parameter extraction and direct channel measurements. Proc. PIMRC 2004 – IEEE 15th Int. Symp. on Pers., Indoor and Mobile Radio Commun, Barcelona, Spain, Sep. 2004. [Also available as TD(04)016.]

[PLVP02] I. Pàez, S. Lorendo, L. Valle, and R. P. Torres. Experimental estimation of wideband radio channel parameters with the use of a spectrum analyzer and the Hilbert transform. *Microwave and Optical Technology Letters*, 43(5): 393–396, Sep. 2002. [Also available as TD(02)074.]

[PPYF04] T. Pedersen, C. Pedersen, X. Yin, B. Fleury, R. Pedersen, B. Bozinovska, A. Hviid, P. Jourdan, and A. Stucki. Investigations of the ambiguity effect in the estimation of Doppler frequency and directions in channel sounding using switched Tx and Rx arrays. *COST 273*, TD(04)021, Athens, Greece, Jan. 2004.

http://www.elektrobit.ch/products/propsound/index.html

[PrSh97] D. T. Prescott and N. V. Shuley. Reflection analysis of FDTD boundary conditions -part I: time-space absorbing boundaries. *IEEE Trans. Microwave Theory Tech.*, 45: 1162–1170, 1997.

[RaAr95] O. M. Ramahi and B. Archambeault. Adaptive absorbing boundary conditions in finite-difference time domain applications for EMC simulations. *IEEE Trans., Electromagn.,Compat.*, 37: 580–583, 1995.

[Rama99] O. M. Ramahi. Stability of absorbing boundary conditions. *IEEE Trans. Antennas Propagat.*, AP-47: 593–599, 1999.

[Rapp86] T. S. Rappaport. *Wireless Communication – Principles and Practice*. Prentice Hall, Upper Saddle River, NJ, USA, 1986.

[RaSa04] N. Razavi-Ghods and S. Salous. Semi-sequential MIMO radio channel sounding. In *Proc. of CCCT04-Int. Conf. on Computing, Communications and Control Technologies*, Austin, TX, USA, Aug. 2004. [Also available as TD(04)079.]

[Rich05] A. Richter. *Estimation of Radio Channel Parameters: Models and Algorithms*. PhD thesis, Ilmenau University of Technology, Ilmenau, Germany, 2005. [Also available as TD(02) 132.]

[RiEK05] A. Richter, M. Enescu, and V. Koivunen. A state space approach to propagation path parameter estimation and tracking. *COST 273*, TD(05)053, Bologna, Italy, Jan. 2005.

[RiOK04] C. Ribeiro, E. Ollila, and V. Koivunen. Stochastic ML method for propagation parameter estimation using mixture of angular distribution models. In *Proc. PIMRC 2004-IEEE 15th Int. Symp. on Pers., Indoor and Mobile Radio Commun.*, Barcelona, Spain, Sep. 2004. [Also available as TD(05)024.]

[RiWG97] K. Rizk, J. F. Wagen, and F. Gardiol. Two-dimensional ray tracing modeling for propagation prediction in micro-cellular environments. *IEEE Trans. Veh. Technol.*, 46(2): 508–517, May 1997.

[RoMGPJL04] J.-V. Rodríguez, J.-M. Molina-García-Pardo, and L. Juan-Llácer. A multiple-building diffraction attenuation function expressed in terms of UTD coefficients for microcellular communications. *Microwave and Optical Technology Lett.*, 40(4): 298–300, Feb. 2004. [Also available as TD(03)153.]

http://www.channelsounder.de/

[SaGo01] S. Salous and H. Gokalp. Characterisation of W-CDMA channels in FDD UMTS bands. In *Proc. VTC 2001 Fall-IEEE 54th Vehicular Technology Conf.*, Atlantic City, NJ, USA, Oct. 2001. [Also available as TD-02–003.]

[SaHi03] S. Salous and V. Hinostroza. Indoor and between building measurements with high resolution channel sounder. In *Proc. Int. Conf. Antennas Propagation*, Exeter, UK, Mar. 2003. [Also available as TD-02–064.]

[SaTA02] K. Sakaguchi, J. Takada, and K. Araki. A novel architecture for MIMO spatio-temporal channel sounder. *IEICE Transactions on Electronics*, E-86C(3): 436–441, Mar. 2002.

[SBAP02] T. C. W. Schenk, R. J. C. Bultitude, L. M. Augustin, R. H. van Poppel, and G. Brussaard. Analysis of propagation loss in urban microcells at 1.9 GHz and 5.8 GHz. In *Proc. of URSI Commission F Open Symposium Radiowave Prop. Remote Sensing*, Feb. 2002. [Also available as TD-03-015.]

[ScFW02] T. Schäfer, T. Fügen, and W. Wiesbeck. Measurement and analysis of radio wave propagation in hospitals. In *Proc. VTC 2002 Fall – IEEE 56th Vehicular Technology Conf.*, Vancouver, Canada, Sep. 2002. [Also availabe as TD(02)056.]

[Schä03] T. M. Schäfer. *Experimental and Simulative Analysis of the Radio Wave Propagation in Hospitals*. PhD thesis, University of

Karlsruhe (TH), Karlsruhe, Germany, 2003. [Also available as TD(02)193.]

[ScLa03] S. Schmidt and G. Lazzi. Extension and validation of a perfectly matched layer formulation for the unconditionally stable D-H FDTD method. *IEEE Microwave Wireless Comp. Lett.*, 13: 345–347, 2003.

[ScWi03] T. M. Schäfer and W. Wiesbeck. Effective modeling of composite walls in hospitals for ray-optical wave propagation simulations. In *Proc. VTC2003 Fall-IEEE 58th Vehicular Technology Conf.*, Orlando, FL, USA, Oct. 2003. [Also available as TD(03)092.]

[SeKa01] S. Semmelrodt and R. Kattenbach. Application of spectral estimation techniques to 2-D fading forecast of time-variant channels. COST 273, TD(01)034, Bologna, Italy, Oct. 2001.

[SeKa02a] S. Semmelrodt and R. Kattenbach. A 2-D fading forecast of time-variant channels based on parametric modeling techniques. In *Proc. PIMRC 2002–IEEE 13th Int. Symp. onPers., Indoor and Mobile Radio Commun.*, Lisbon, Portugal, Sep. 2002.

[SeKa02b] S. Semmelrodt and R. Kattenbach. Efficient implementation of an extended SAGE algorithm for the extraction of time-varying channel parameters. *COST 273*, TD(02)120, Lisbon, Portugal, Sep. 2002.

[SeKa03] S. Semmelrodt and R. Kattenbach. Investigation of different fading forecast schemes for flat fading radio channels. In *Proc. VTC 2003 Fall – IEEE 58th Vehicular Technology Conf.*, Orlando, FL, USA, Oct. 2003. [Also available as TD-03-045.]

[Semm04a] S. Semmelrodt. *Methoden zur prädiktiven Kanalschätzung für adaptive Übertragungstechniken im Mobilfunk*. PhD thesis, Universität Kassel, Kassel, Germany, 2004. In German, ISBN 3-89958-041-9.

[Semm04b] S. Semmelrodt. Spectral analysis and linear prediction toolbox for stationary and non-stationary signals. *FREQUENZ*, 58(7–8): 185–187, 2004. [Also available as TD(04)019.]

[SFIH02] S. Salous, P. Fillipides, and I. Hawkins. Multiple antenna channel sounder using a parallel receiver architecture. In *Proc. of SCI'02 – 6th World Multi-Conf. on Systemics, Cybernetics and Informatics*, Orlando, FL, USA, July 2002. [Also available as TD(02)002.]

[SHBPM03] S. W. Staker, C. L. Holloway, A. U. Bhobe, and M. Piket-May. Alternating-direction implicit (ADI) formulation of the finite-difference time-domain (FDTD) method: algorithm and material dispersion implementation. *IEEE Trans. Electromagn. Compat.*, 45: 156–166, 2003.

[SICW01] H. Shimotahira, K. Iizuka, S. Chu, C. Wah, F. Costen, and S. Yoshikuni. Three-dimensional laser microvision. *Applied Optics*, 40(11): 1784–1794, 2001. [Also available as TD(02)007.]

[SiMi03] I. Sirkova and M. Mikhalev. Influence of tropospheric ducting on microwave propagation in short distances. In *Proc. of ICEST2003, 38th International Conference on Information, Communication and Energy Systems and Technologies*, Sofia, Bulgaria, Oct. 2003. [Also available as TD(02)086 and TD(02)152.]

[SiMi04] I. Sirkova and M. Mikhalev. Parabolic equation based study of ducting effects on microwave propagation. *Microwave and Optical Technology Lett.*, 42(5): 390–394, Sep. 2004. [Also available as TD(02)152 and TD(05)006.]

[SiMi05] I. Sirkova and M. Mikhalev. Digital terrain elevation data combined with the PE method: a Sofia region study. *COST 273*, TD(05)005, Bologna, Italy, Jan. 2005.

[SkCK05] N. Skentos, P. Constantinou, and A. G. Kanatas. Results from rooftop to rooftop MIMO channel measurements at 5.2 GHz. *COST 273*, TD(05)059, Bologna, Italy, Jan. 2005.

[SKPC04] N. Skentos, A. G. Kanatas, G. Pantos, and P. Constantinou. Channel characterization results from fixed outdoor MIMO measurements. In *Proc. WPMC 2004–Wireless Pers. Multimedia Commun.*, Abano Terme, Italy, Sep. 2004. [Also available as TD-04-140.]

[SLRT03] G. Sommerkorn, M. Landmann, R. Richter, and R. Thomä. Performance evaluation of real antenna arrays for high-resolution DoA estimation in channel sounding–part 2: experimental ULA measurement results. *COST 273*, TD(03)196, Sep. 2003.

[Smit65] G. D. Smith. *Numerical Solution of Partial Differential Equations*. Oxford University Press, Oxford, UK, 1965.

[StFJ02] A. Stucki, B. Fleury, and P. Jourdan. Isis, a high performance and efficient implementation of SAGE for radio channel parameter estimation. *COST 273*, TD(02)068, Espoo, Finland, May 2002.

[StMo97] P. Stoica and R. Moses. *Introduction to Spectral Analysis*. Prentice-Hall, Upper Saddle River, NY, USA, 1997.

[TaBN04] C. Tan, M. Beach, and A. R. Nix. Multipath parameters estimation with a reduced complexity unitary-SAGE algorithm. *European Transactions on Communications*, (14): 515–528, Jan. 2004. [Also available as TD(03)090.]

[Tafl95] A. Taflove. *Computational Electrodynamics*. Artech House, Norwood, MA, 1995.

[Taga97] T. Taga. Smart MUSIC algorithm for DOA estimation. *Elect. Lett.*, 33(3): 190, 1997. [Also available as TD(01)001.]

[TaLH02] R. Tahri, C. Letrou, and F. V. Hanna. A beam launching method for propagation modeling in multipath contexts. *Microwave and Optical Technology Lett.*, 35(1):6–10, Oct. 2002. [Also available as TD(02)031.]

[TaSh96] F. Taga and H. Shimitahira. Proposal of the fast kernel MUSIC algorithm. *IEICE Trans. Fundamentals*, E79-A(8):1232, 1996. [Also available as TD(01)016.]

[Teal01] P. D. Teal. *Real Time Characterisation of the Mobile Multipath Channel*. PhD thesis,

University of Sydney, Sydney, Australia, 2001.

[TFBN04] C. M. Tan, S. E. Foo, M. A. Beach, and A. R. Nix. Descriptions of dynamic single-, double-directional measurement campaigns at 5 GHz. *COST 273*, TD-04–099, June 2004.

[ThCo04] A. Thiry and F. Costen. On the implementation of the frequency-dependent alternating direction-implicit finite difference time domain method. In *Proc. of IEE Seminar on Ultra Wideband Communications Technologies and System Design: Technical and Operational Development*, 2004.

[ThLR04] R. Thomä, M. Landmann, and A. Richter. RIMAX a maximum likelihood frame-work for parameter estimation in multidimensional channel sounding. In *Intl. Symp. on Antennas and Propagation*, Sendai, Japan, Aug. 2004. [Also available as TD(04)045.]

[ThYo03] J. S. Thompson and S. K. Yong. A closed-form spatial fading correlation model for electromagnetic vector sensors. *COST 273*, TD(03)106, Paris, France, May 2003.

[TLAP02] C. M. Tan, M. Landmann, A. Richter, L. Pesik, M. A. Beach, C. Schneider, R. S. Thomä, and A. R. Nix, On the application of circular arrays in direction finding part II: experimental evaluation on SAGE with different circular arrays. In *1st Annual COST 273 Workshop*, Espoo, Finland, May 2002.

[TLRT05a] R. Thomä, M. Landmann, A. Richter, and U. Trautwein. *Multidimensional High Resolution Channel Sounding, in Smart Antennas in Europe, State-of-the-Art*. EURASIP Book Series, 2005. [Also available as TD(03)198.]

[TLRT05b] R. Thomä, M. Landmann, A. Richter, and U. Trautwein. *Smart Antennas in Europe – State-of-the-Art*. Hindawi, Sylvania, OH, USA, 2005. [Also available asTD-03-198.]

[TLST05] U. Trautwein, M. Landmann, G. Sommerkorn, and R. Thomä. System-oriented measurement and analysis of MIMO channels. *COST 273*, T(05)063, Bologna, Italy, Jan. 2005.

[TLVD01] R. P. Torres, S. Loredo, L. Valle, and M. Domingo. An accurate and efficient method based on ray-tracing for the prediction of local flat-fading statistics in pic-ocell radio channels. *IEEE J. Select. Areas Commun.*, 19(2): 170–178, Feb. 2001. [Also available as TD(01)014.]

[ToBL98] S. A. Torrico, H. L. Bertoni, and R. H. Lang. Modelling tree effects on path loss in a residential environment. *IEEE Trans. Antennas Propagat.*, 46: 872–880, 1998.

[ToLa04] S. Torrico and R. Lang. Total attenuation through a two-dimensional trunk dominated forest. In *Proc. of National URSI Meeting*, Boulder, CO, USA, Jan. 2004. [Also available as TD(04)105.]

[TrLC04] D. Trappeniers, E. Van Lil, and A. Van de Capelle. Cylindrical and spherical obstacles in epics-GO. *Revue HF Tijdschrift; Belgian Journal of Electronics and Communications*, 2004(2): 20, Apr. 2004. [Also available in extended version as TD(03)104.]

[TrSi02] V. P. Tran and A. Sibille. Inter-sensor coupling and spatial correlation effects on the capacity of compact MIMO antennas. *COST 273*, TD(02)128, Lisbon, Portugal, Sep. 2002.

[TrSi03] V. P. Tran and A. Sibille. MIMO channel capacity and mutual coupling in circular arrays of monopoles. *COST 273*, TD(03)099, Paris, France, Jan. 2003.

[TsHT04] H. Tsuchiya, K. Haneda, and J. Takada. UWB indoor double-directional channel sounding for understanding the microscopic propagation mechanisms. In *Proc. WPMC 2004 – Wireless Pers. Multimedia Commun.*, Padova, Italy, Sep. 2004. [Also available as TD(04)192.]

[UlMF96] F. T. Ulaby, R. K. Moore, and A. K. Fung. *Matrix Computations*. Johns Hopkins, Baltimore, MD, USA, 1996.

[VeLC02a] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Influence of coupling in antenna arrays on the SAGE algorithm. In *Proc. of XXVII URSI General Assembly*, Maastricht, The Netherlands, Aug. 2002. [Also available as TD(02)060.]

[VeLC02b] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Study and implementation of a uniform spherical distributed antenna array. In *Proc. of URSI Open Symp. on Propagation and Remote Sensing*, Garmisch-Partenkirchen, Germany, Feb. 2002. [Also available as TD(02)1 16.]

[VeLC03a] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Experimental and theoretical direction of arrival extraction with a Mill's cross antenna array. In *Industry-Ready Innovative Research, 1st Flanders Engineering PhD Symp.*, Brussels, Belgium, Dec. 2003. [Also available as TD(03)158.]

[VeLC03b] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Extraction of source parameters from broadband measurements with the SAGE algorithm. *Revue HF, Belgian Journal of Electronics and Communications*, (3):41–43, 2003. [Also available as TD(03) 108.]

[VeLC04a] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Direction of arrival (DOA) parameter estimation with the SAGE algorithm. *Elsevier Signal Processing*, 84(3):619–629, Mar. 2004. [Also available as TD(02)020.]

[VeLC04b] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Verification of the BTD-SAGE algorithm with simulated and experimental data. *COST 273*, TD(04)022, Athens, Greece, Jan. 2004.

[VeLC05] J. Verhaevert, E. Van Lil, and A. Van de Capelle. Applications of the SAGE algorithm using a dodecahedral receiving antenna array. *COST 273*, TD(05)013, Bologna, Italy, Jan. 2005.

[Vite94] A. J. Viterbi. Other-cell interference in cellular power-controlled CDMA. *IEEE Trans. Commun.*, 42(2): 1501–1504, Feb. 1994.

[VLSK03] J. Verhaevert, E. Van Lil, S. Semmelrodt, R. Kattenbach,

and A. Van de Capelle. Analysis of the SAGE DOA parameter extraction sensitivity with 1.8 GHz indoor measurements. In *Proc. VTC 2003 Fall–IEEE 58th Vehicular Technology Conf.*, Orlando, FL, USA, Oct. 2003. [Also available as TD(03)035.]

[WaCh95] R. L. Wagner and W. C. Chew. An analysis of Liao's absorbing boundary conditions. *J. Electromagn. Waves Applicat.*, 9:993–1009, 1995.

[WaHo03] A. Waern and P. Holm. Wave propagation over a forest edge parabolic equation modelling vs. GTD modelling. *COST 273*, TD(03)169, Prague, Czech Republic, Sep. 2003.

[WAKE04] S. Wyne, P. Almers, J. Karedal, G. Ericsson, F. Tufvesson, and A. F. Molisch. Outdoor to indoor office MIMO measurements at 5.2 GHz. In *Proc. VTC 2004 Fall-IEEE 60th Vehicular Technology Conf.*, Los Angeles, CA, USA, Sep. 2004. [Also available as TD(04)152.]

[WHZL02] G. Woelfle, R. Hoppe, D. Zimmermann, and F. M. Landstorfer. Enhanced localization technique within urban and indoor environments based on accurate and fast propagation models. In *Proc. WIRELESS 2002–Proc. 15th Int. Conf. on Wireless Commun.*, Florence, Italy, Feb. 2002. [Also available as TD(02)033.]

[WLGB97] G. Woelfle, F. M. Landstorfer, R. Gahleitner, and E. Bonek. Extensions to the field strength prediction technique based on dominant paths between transmitter and receiver in indoor wireless communications. In *Proc. EPMCC 1997-2nd European Personal and Mobile Communications Conference*, Bonn, Germany, Sep. 1997.

[WoLa98] G. Woelfle and F. M. Landstorfer. Dominant paths for the field strength prediction. In *Proc. (VTC) 1998–(IEEE) 48th Vehicular Technology Conf.*, Ottawa, Ontario, Canada, May 1998.

[WoZo00] K. T. Wong and M. D. Zoltowski. Closed-form direction finding and polarization estimation with arbitrarly spaced electromagnetic vector-sensors at unknown locations. *IEEE Trans. Antennas Propagat.*, 48(5): 671–681, May 2000.

[WWWW04] G. Woelfle, P. Wertz, R. Wahl, P. Wildbolz, and F. M. Landstorfer. Dominant path prediction model for indoor and urban scenarios. *COST 273*, TD(04)205, Duisburg, Germany, Sep. 2004.

[WZWH03] P. Wertz, D. Zimmermann, G. Woelfle, R. Hoppe, and F. M. Landstorfer. Hybrid ray optical models for the penetration of radio waves into enclosed spaces. In *Proc. VTC 2003 Fall – IEEE 58th Vehicular Technology Conf.*, Orlando, FL, USA, Oct. 2003. [Also available as TD(03)177.]

[XaVC03] D. M. Xavier, J. M. Venes, and L. M. Correia. Characterisation of signal penetration into buildings for GSM. In *Proc. of Conf Tele 2003-4th Conference on Telecommunications*, Aveiro, Portugal, June 2003. [Also available as TD(03)069.]

[Yee66] K. S. Yee. Numerical solution of initial boundary value problems involving Maxwell's equations in isotropic media. *IEEE Trans. Antennas Propagat.*, AP-14, 1966.

[YiFS03] X. Yin, B. Fleury, and A. Stucki. Doppler estimation for channel sounding using switched multiple transmit and receive antennas. *COST 273*, TD(03)026, Paris, France, May 2003.

[YuCh03] C. Yuan and Z. Chen. On the modeling of conducting media with the unconditionally stable ADI-FDTD method. *IEEE Trans. Microwave Theory Tech.*, 51-8:1929–1938, 2003.

[ZBLL04] D. Zimmermann, J. Baumann, M. Layh, F. M. Landstorfer, R. Hoppe, and G. Woelfle. Database correlation for positioning of mobile terminals in cellular networks using wave propagation models. In *Proc. VTC 2004 Fall–IEEE 60th Vehicular Technology Conf.*, Los Angeles, CA, USA, Sep. 2004. [Also available as TD(04)195.]

[ZeTS04] R. Zetik, R. Thomä, and J. Sachs. Ultra-wideband real-time channel sounder and directional channel Parameter estimation. In *Proc. of URSI Int. Symp. on Electromagnetic Theory*, Pisa, Italy, May 2004. [Also available as TD(03)201.]

[ZhCZ00] F. Zheng, Z. Chen, and J. Zhang. Toward the development of a three-dimensional unconditionally stable finite-difference time-domain method. *IEEE Trans. Microwave Theory Tech.*, 48:1550–1558, 2000.

[ZVRK04] X. Zhao, P. Vainikainen, T. Rautiainen, and K. Kalliola. Path loss models for urban microcells at 5.3 GHz. *COST 273*, TD(04)207, Duisburg, Germany, Sep. 2004.

This page intentionally left blank

Section **Five**

Short range wireless communication

This page intentionally left blank

Chapter 5.1

5.1

Wireless local area networks

Daniel Dobkin

5.1.1 Networks large and small

Although this book is primarily about voltages and waves, to make the text reasonably self-contained, we provide a brief introduction into the topic of networking in general and review the characteristics of several wireless local network technologies. In this review, the reader will encounter real examples of how the common concepts are put to use in getting data from one place to another wirelessly.

Our treatment is necessarily cursory and emphasizes aspects relevant to our main topic of wireless transmission and reception, giving short shrift to frame formats, management entities, and many other topics directed toward network implementation.

Data networks have existed in various forms since the invention of telegraphy in the nineteenth century. However, the use of data networks has expanded tremendously with the proliferation of computers and other digital devices. Today, networks extend everywhere, and the Internet—the network of networks—allows computers almost anywhere in the world to communicate with each other using these networks.

Data networks can be organized in a hierarchy based on physical extent. The physical reach of the network is a useful figure of merit both for technical and practical reasons. Information cannot move faster than the speed of light, so as the distance covered by a network grows, its *latency*—the minimum time needed to send a bit between stations—grows too. Latency has important consequences on how networks are designed and operated: if the transmitter has to wait a long time for a reply to a message, it makes sense to send large chunks (*packets*) of information at once, because otherwise the medium will simply be idle while awaiting confirmation. Long links also need low error rates, because the time required to correct an error is long. Shorter links can take a different approach, sending a packet and then waiting for a response, because the time required is not large. Small and large networks are also practically and commercially distinct: small networks are usually owned and operated by their users, whereas large networks are owned and operated by service providing companies who are not the primary users of their capacity.

With that brief preface, let us categorize data networks by size:

- *Personal area networks* (PANs): PANs extend a few meters and connect adjacent devices together. To speak formally, a data connection between a single transmitting station and a single receiving station is a link, not a network, so the connection between, for example, a desktop computer and an external modem (if there are any readers who remember such an impractical arrangement) is a data *link* rather than a data network. However, more sophisticated cabling systems such as the small computer system interface, which allow sharing of a single continuous cable bus between multiple stations, may be regarded as networks. More recently, wireless PANs (WPANs) have become available, with the ambition to replace the tangle of cabling that moves data between devices today. Wired PANs were purely dedicated to moving data, though some WPAN technologies (such as Bluetooth, discussed in section 5.1.5.1) also support voice traffic.
- *Local area networks* (LANs): LANs were invented to connect computers within a single facility, originally defined as a room or small building and later extended to larger facilities and multisite campuses. LANs extend a few hundred meters to a few

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2005 Elsevier Inc. All rights reserved

kilometers. They are generally owned and operated by the same folks who own the site: private companies, individuals in their homes, government agencies, and so on. They are generally used indoors and historically have been solely for the movement of data, though the recent implementation of voice-over-Internet-protocol technologies has allowed LANs to provide voice service as well. Ethernet (see section 5.1.2) is by far the most prevalent LAN technology.

- *Metropolitan area networks* (MANs): MANs connect different buildings and facilities within a city or populated region together. There is a significant technological and historical discontinuity between LANs and MANs: LANs were invented by the computer-using community for data transfer, whereas MANs descended primarily from the telephone network, traditionally organized to move time-synchronous voice traffic. MANs are generally owned by local telephone exchanges (incumbent local exchange companies) or their competitors (competitive local exchange companies). They are organized around a large number of feeders to a small number of telephone central offices, where traffic is aggregated. Most MANs deployed today are based on a hierarchical system of increasingly faster synchronous data technologies. T-1 lines (in the United States) provide 1.5 megabits per second (Mbps) of data over twisted pairs of copper wires and in sheer numbers still dominate over all other connections to the MAN: there are thousands or tens of thousands of these in a large central office. T-1 and T-3 (45 Mbps) are further aggregated into faster connections, usually using fiber optic transmission over synchronous optical network links: OC-3 at 155 Mbps, OC-12 at 622 Mbps, and so on. Traditional MANs support both voice and data transfer.
- *Wide area networks* (WANs): WANs connect cities and countries together. They are descended from the long-distance telephone services developed in the mid-twentieth century and are generally owned and operated by the descendants of long-distance telephone providers or their competitors, where present. Almost all long-distance telecommunication today is carried over fiber-optic cables, commonly at OC-12 (622 Mbps), OC-48 (2.5 Gbps), or OC-192 (10 Gbps) rates; OC-768 (40 Gbps) is in the early stages of deployment in 2004. WAN connections cross the oceans and continents and carry the voice and data commerce of the world. The Internet initially evolved separately from WANs, but its explosive growth in the late 1990s resulted in a complex comingling of the traditional WAN and LAN communities, businesses, and standards bodies.

In the simplest view, there are small networks (PANs and LANs) and big networks (MANs and WANs). Small networks deliver best-effort services over short distances and are cost sensitive. Big networks deliver guaranteed-reliable services over long distances and are quality sensitive and cost competitive. This book focuses on how small networks are converted from wires to wireless links.

All networks, big or small, have certain elements in common. A message (typically a packet in a data network) must be addressed to the destination station. The message must have a format the destination station can understand. The message has to get access to some physical medium to be sent. Errors in transmission must be corrected. These activities can be grouped into a hierarchical arrangement that helps provide structure for designing and operating the networks. *Applications* live on the top of the hierarchy and use the networking services. *Networks* support applications and deal with routing messages to the destination station. *Links* transfer data between one station and another in the network and are responsible for access to the physical medium (a wire or a radio link), packaging the packet in an appropriate format readable by the receiving station and then reading received packets and producing the actual voltages or signals. (Note that an important function of networks, *correcting errors* in messages, can go anywhere in the hierarchy and may be going on simultaneously at some or all of the levels, leading on occasion to regrettable results.)

A very widely used example of this sort of hierarchical arrangement is the *Open Systems Interconnect* (OSI) protocol stack. A simplified view of the OSI stack, arranged assuming a wireless LAN (WLAN) link to one of the stations, is depicted in Figure 5.1-1. The arrangement is in general correspondence with the requirements noted above, though additional layers have been interposed to provide all the manifold functions needed in complex networks.

The standards work of the Institute of Electrical and Electronics Engineers (IEEE), with which we shall be greatly concerned in this chapter, generally divides the data link layer into an upper *logical link control* layer and a lower *medium access control* (MAC) layer. In this chapter, we are almost exclusively interested in what goes on in the link layer and the physical (PHY) layer. The logical link control layer is focused mainly on the requirements of the higher layers, and in practice the logical link control layer used for many WLAN technologies is the same as the one used for wired technologies. We do not discuss it further. However, the MAC layer is intimately involved in the efficient use of the wireless medium, and it is of great interest for any user of a wireless link.

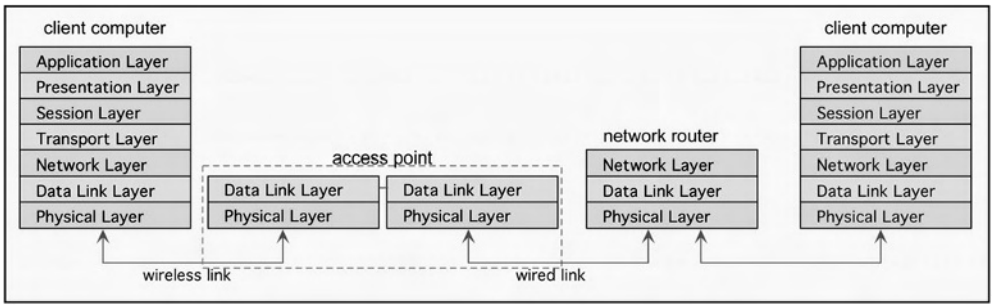


Figure 5.1-1 The OSI protocol stack; the configuration shown is typical of an 802.11 infrastructure network.

5.1.2 WLANs from LANs

LANs connect nearby computers together so that they can exchange data. Ethernet, by far the most popular LAN technology today, was invented by Bob Metcalfe of Xerox Palo Alto Research Center in 1973, in connection with the prescient work on personal computers then being done at Palo Alto Research Center. Ethernet is an *asynchronous* technology: there are no fixed time slots assigned to users. There is also no central control of access to the transmitting medium. Instead, Ethernet exploits three basic ideas to minimize the impact of two or more stations trying to use the medium at the same time; two of them depend on the ability of a station to listen to the voltages on the cable while it sends its message. C*arrier sensing* is used by each station to make sure no one is using the medium before beginning a transmission. (Note that in fact the most popular forms of Ethernet use baseband signaling, that is, the voltages on the cable correspond directly to binary bits and are not modulating a high-frequency carrier, so there's no carrier to sense. Why did you expect it to make sense?) C*ollision detection*—noticing that the signal on the cable isn't the one the station is sending—allows the station to detect that another station has tried to send a message at the same time it is transmitting. (The combination of these two techniques is often abbreviated C*SMA/CD*, for carrier-sense multiple access with collision detection.) Finally, *random backoff* is used after a collision to ensure that stations won't simply continue to collide: after all stations become silent, each station that has a pending packet to send randomly chooses a time delay before attempting to send it. If another collision results, the stations do the same exercise, but over a larger time window. This approach has the effect of forcing the *offered traffic* to the network to go down when collisions are frequent, so that data transfer continues at a lower rate but the network doesn't get tangled up: ethernet networks fail gracefully when they get busy. Thus, Ethernet was an early example of a MAC layer designed to allow peer stations to share a medium without central coordination.

Wireless stations are mobile, and access to the wireless medium cannot be as readily constrained as a wired connection. Wireless links are much noisier and less reliable in general than wired links. The MAC and PHY layers of a wireless network must therefore deal with a number of issues that are rare or absent from the provision of a wired link:

- *Getting connected*: Wireless stations by default are likely to be mobile. How does a wireless station let other mobile or fixed stations know it is present and wants to join the network? This is the problem of *associating* with the network. A closely related problem is the need of portable mobile stations to save power by shutting down when they have no traffic, without losing their associated status; the network coordinator must remember who is awake and who is asleep, allow new stations to join, and figure out that a station has left.
- *Authentication*: In wired networks, a station plugs into a cable placed by the network administrator. If you've entered the building, you're presumed to be an authorized user. Wireless propagation is not so well controlled, so it is important for the network to ensure that a station ought to be allowed access and equally important for the station to ensure that it is connecting to the intended network and not an impostor.
- *Medium access control*: The wireless medium is shared by all (local) users. Who gets to transmit what when? A good wireless network protocol must efficiently multiplex the shared medium.
- *Security*: Even if the network refuses to talk to stations it doesn't recognize, they might still be listening in. Wireless networks may provide additional security for the data they carry by *encryption* of the data stream. It is important to note that security may also be provided by other layers of the network protocol stack; the lack of local encryption in the wireless link does not necessarily imply insecurity, and encryption of the wireless data

doesn't ensure security from eavesdropping at some other point in the network.

- *Error correction*: The wireless medium is complex and time varying. There is no way to transmit all packets without errors. The link layer may insist on receiving a positive *acknowledgment* (ACK) before a station may consider a transmission successful. Incoming data packets may be *fragmented* into smaller chunks if packet errors are high; for example, in the presence of microwave oven interference, which peaks in synchrony with power lines at 60 Hz, fragmentation may be the only way to successfully transmit very long packets. Fragmented packets must be identified as such and reassembled at the receiving station.
- C*oding and interleaving*: Errors are inevitable in the wireless medium. Protection against bit errors can be provided by encoding the data, so that errors are detected by the receiving station. Codes often allow errors of up to a certain size to be corrected by the receiver without further discussion with the transmitting station: this is known as *forward error correction*. Interleaving is the process of redistributing bits into an effectively random order to guard against bursts of errors or interference destroying all the neighboring bits and thus defeating the ability of the codes to correct local errors.
- *Packet construction*: Data packets get *preambles* prepended onto them. The preambles typically contain synchronization sequences that allow the receiving station to capture the timing of the transmitter and, in the case of more sophisticated modulations, may also allow the receiver to determine the carrier phase and frequency. The preambles also contain digital information specific to the wireless medium.
- *Modulation and demodulation*: The resulting packets must then be modulated onto the carrier, transmitted on the wireless medium, and received, amplified, and converted back into bits.

Let us examine how these problems are addressed in some current and upcoming WLAN and WPAN technologies.

5.1.3 802.11 WLANs

In 1985 the U.S. Federal Communications Commission (FCC) issued new regulations that allowed unlicensed communications use of several bands, including the 2.4-GHz band, that had previously been reserved for unintended emissions from industrial equipment. Interest in possible uses of this band grew, and in the late 1980s researchers at NCR in Holland, who had experience in analog telephone modems, initiated work to develop a wireless data link. The initial experimental units used an existing Ethernet MAC chip. As discussed in more detail in section 5.1.3.2, there is no easy way to detect a collision in a wireless network; the experimenters worked around this limitation by informing the MAC chip that a collision occurred any time a positive ACK was not received, thus initiating the Ethernet backoff algorithm. In this fashion they were able to make working wireless links with minimal modification of the existing Ethernet MAC. The early workers understood that to make a true volume market for these products, standardization was necessary. When standardization efforts were introduced at the IEEE shortly thereafter, the MAC was elaborated to deal with the many wireless-specific issues that don't exist in Ethernet but remained based on Ethernet, which had been introduced because of its availability and was successful because of its simplicity and robustness. Thus, the standard today continues to reflect both the strengths and limitations of the original Ethernet MAC.

The IEEE 802 working group deals with standards for wired LANs and MANs. IEEE 802.3 is the formal standardization of the Ethernet wired LAN. The IEEE decided to incorporate WLANs as part of the 802 working group and created the 802.11 activity, culminating in the first release in 1997 of the 802.11 standard. In recent years, in addition to various elaborations of the 802.11 standards, which are discussed below, other working groups have formed within 802 to consider related applications of wireless data links. (In practice, WPANs are also included as part of the 802.15 activity.)

The 802.11 standard actually allowed three physical layers: an infrared link, a *frequency-hopping* (FH) radio link, and a *direct-sequence spread-spectrum* (DSSS) radio link. These links supported data rates of 1 to 2 Mbps. The infrared link physical layer has had little commercial significance (in fact, the author has been unable to find any evidence of any commercial product using this protocol ever having been manufactured). Commercial products were deployed with the FH radio, and the FH approach does offer certain advantages in terms of the number of independent collocated networks that can be supported with minimal interference. However, in terms of current commercial importance, the DSSS physical layer is completely dominant, due to its later extension to higher data rates in the 802.11b standard in 1999. Therefore, here we shall concentrate only on this physical layer. The IEEE standards bodies do many wonderful things, but creation of convenient nomenclature is not one of them; therefore, we introduce the terminology *802.11 classic* to refer to the original 1997 802.11 specification and its later releases and in our case predominantly to the DSSS variant. We can then unambiguously use the unmodified moniker "802.11" to

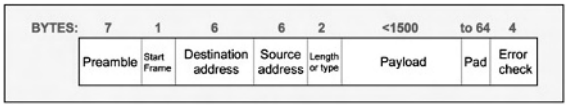


Figure 5.1-2 An Ethernet (802.3) frame (Packet).

refer to all the 802.11 working group standards, including the alphabet soup of elaborations released in later years.

5.1.3.1 802.11 Architecture

The 802.11 standard allows both infrastructure networks, which are connected to a wired network (typically Ethernet) using an *access point*, and independent networks connecting peer computers wirelessly with no wired network present. Most installations are of the infrastructure variety, and we focus on them. From the point of view of the Ethernet network, which is usually connected to both ends of an 802.11 link, the wireless link is just another way of moving an Ethernet packet, formally known as a *frame* (Figure 5.1-2), from one station to another (Figure 5.1-3).

However, the possibility (well, likelihood) that some or most wireless stations are mobile stations means that the architecture of a wired Ethernet network with wireless stations is likely to be fundamentally different from a conventional wired network. The stations associated to a particular access point constitute a *basic service set* (BSS).

The collection of BSSs connected to a single wired network forms an *extended service set* (ESS) (Figure 5.1-4). A BSS has a unique identifier, the *BSSID*. An ESS also has an identifier, the *ESSID*, that is unique to that ESS but shared by all the component BSSs (is that enough acronyms in one sentence for you?). Unfortunately, interfaces for many commercial 802.11 implementations use the nomenclature SSID, which in practice usually refers to the ESSID but is obviously somewhat ambiguous and potentially confusing. The ESS architecture provides a framework in which to deal with the problem of mobile stations *roaming* from one BSS to another in the same ESS. It would obviously be nice if a mobile station could be carried from the coverage area of one access point (i.e., one BSS) to that of another without having to go through a laborious process of reauthenticating, or worse, obtaining a new Internet protocol (IP) address; ideally, data transfers could continue seamlessly as the user moved. A distribution system is assigned responsibility for keeping track of which stations are in which BSSs and routing their packets appropriately. However, the original and enhanced 802.11 standards did not specify how such roaming ought to be conducted; only with the release of the 802.11f *Inter-Access Point Protocol* (IAPP) standard in 2003 did the operation of the distribution system receive a nonproprietary specification. The IAPP protocol uses the nearly universal internet protocol (TCP/IP) combined with RADIUS servers to provide secure communication between access points, so that moving clients can reassociate and expect to receive forwarded packets from the distribution system. In the future, local roaming on 802.11 networks should be (hopefully) transparent to the user, if compliant access points and network configurations become widely established.

5.1.3.2 MAC and CSMA/CA

Despite its name, Ethernet was conceived as a cable-based technology. Putting all the stations on a cable and limiting the cable length allowed meant that all stations could always hear one another, and any collision would be detected by all the stations on the cable. However, in a wireless network it is unlikely that all stations will be

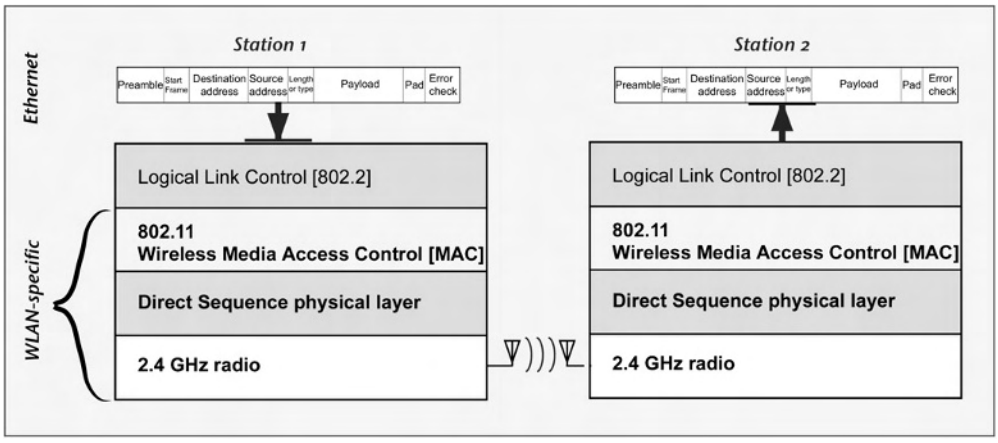


Figure 5.1-3 802.11 as a means of moving Ethernet frames.

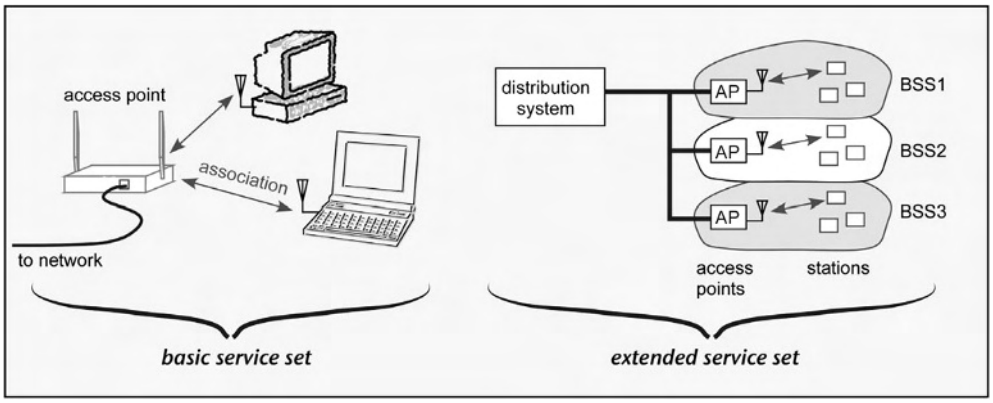


Figure 5.1-4 Definition of the BSS and ESS.

able to receive all transmissions all the time: this is known as the *hidden station* problem. An example is depicted in Figure 5.1-5: the access point can communicate with stations A and B, but A and B cannot receive each other's transmissions directly. Because of this fact, collision detection would fail if A and B both attempted to transmit at the same time. Furthermore, as we've already seen, received signals are tiny. Provisions to separate a tiny received signal from a large transmitted signal are possible if the two are at differing frequencies, but this requires bandwidth that was not available in the limited 2.4-GHz ISM band. It is difficult and relatively expensive to reliably receive a tiny transmitted signal while simultaneously transmitting at the same frequency, so collision detection is not practical for a WLAN station. Thus, in a wireless implementation, CSMA/CD cannot be used, because there is no way to confidently associate the absence of a carrier with a free medium or reliably detect all collisions.

To get around this difficulty, 802.11 stations use *carrier-sense multiple access with collision avoidance* (CSMA/CA, Figure 5.1-6). All stations listen to the current radio channel before transmitting; if a signal is detected, the medium is considered to be busy and the station defers its transmission. A virtual carrier sense mechanism—the *network allocation vector* (NAV)—is provided to further reduce the likelihood of collisions. Each packet header involved in an exchange that lasts longer than a single frame, such as a data frame that expects to be acknowledged by the receiver, will provide a NAV value in the header of the frame. All stations that receive the frame note the value of the NAV and defer for the additional required time even if they can't detect any signal during that time.

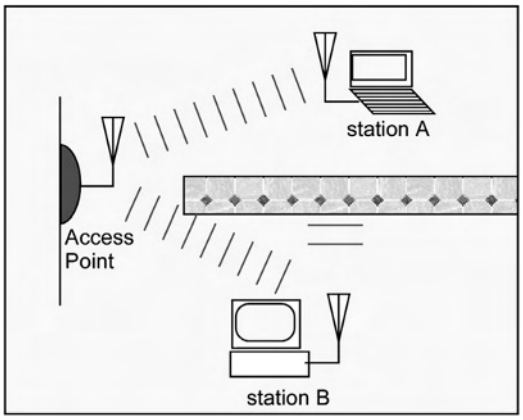


Figure 5.1-5 Stations A and B are hidden from one another, though both are visible to the access point.

Requirements on the timing between frames—interframe spaces—are used to enable flexible access control of the medium using the carrier sense mechanisms. Once a transmission has begun, the transmitting station has captured the medium from all other stations that can hear its transmissions using the physical carrier sense mechanism. A station wishing to send a new packet must first ensure that both the physical carrier sensing mechanism and the NAV indicate the medium to be free for a time equal to the *distributed interframe space* (*DIFS*). The distributed interframe space is relatively long, so that stations that are completing exchanges already in progress, which are permitted to transmit after a *short interframe space* (*SIFS*), can capture the medium to transmit their response packets before any new station is allowed to contend for the right to transmit. The NAV value of each packet in a sequence is also set to capture the medium for the entire remaining sequence of packets. Once all the necessary parts of an exchange are complete, the medium may then become free for the distributed interframe space, at which point stations waiting to transmit frames randomly select a time slot and start transmission if no carrier has been detected.

Because collisions cannot be detected, the opposite approach is taken: 802.11 depends on positive acknowledgement of a successful transmission through the receipt of an ACK packet by the transmitting station. Failure to receive an ACK, whether due to poor signal strength, collisions with other stations, or interference, is

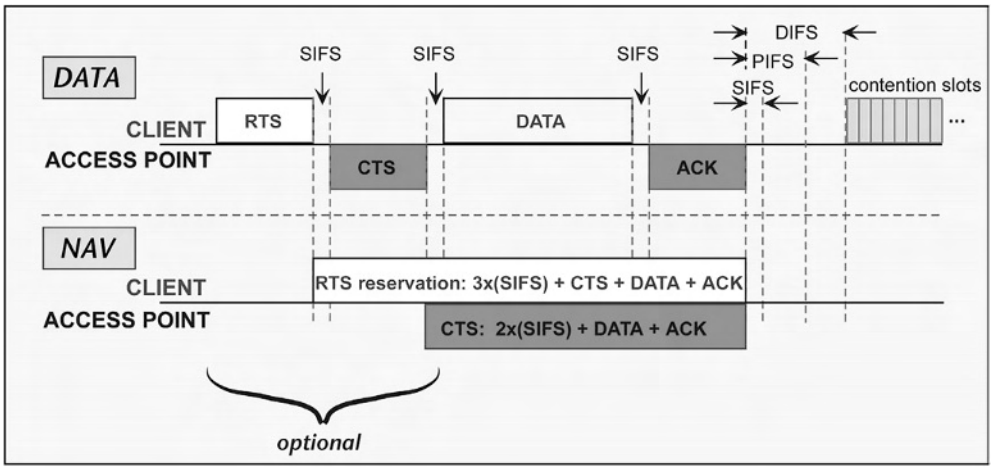


Figure 5.1-6 Example of packet exchange under CSMA/CA.

considered to be indicative of a collision; the transmitting station will wait until the medium is again free, choose a random time slot within a larger possible window (the Ethernet backoff mechanism), and attempt to transmit the packet again.

Two additional optional provisions can be taken to improve performance when, for whatever reason, packets are lost frequently. First, the sending station (typically a client rather than an access point) can precede its data transmission with a *request to send* (RTS) packet. This sort of packet informs the access point and any other stations in range that a station would like to send a packet and provides the length of the packet to be sent. The access point responds after the short interframe space with a *clear to send* (CTS) packet, whose NAV reserves the medium for the time required for the remainder of the data transmission and ACK. The advantage of the scheme is that one can hope that all associated stations can hear the access point CTS packet even if they are not able to detect the RTS packet. An exchange like this is depicted schematically in Figure 5.1-6. The exchange begins with the client's RTS packet, which reserves the NAV for the whole of the envisioned exchange (the length of the data packet, three short interframe spaces, a CTS frame, and an ACK). The access point responds with clearance, reserving the medium for the remainder of the exchange with its NAV setting. All stations then defer until the end of the exchange. After a distributed interframe space has passed without any more transmissions, a station waiting to send data can use one of the contention slots and transmit.

The second backup mechanism for improving the chances of getting a packet through is *fragmentation*. A large packet from the wired medium can be split into smaller fragments, each of which is more likely to be transmitted without a problem. If an error is encountered in one of the fragments, only that fragment needs to be resent. Like the RTS/CTS exchange, a station transmitting a packet sets the NAV value to reserve the medium for the whole exchange consisting of all the fragments of the packet and an ACK for each fragment.

An important consequence of the MAC is that an 802.11 radio is never transmitting and receiving simultaneously: it is a *half-duplex* system in radio terms. This choice simplifies the design of the radio front end by eliminating the need to distinguish between the powerful transmitted signal and a tiny received signal.

The MAC layer defined by the original 802.11 standard is used almost unchanged for the enhanced versions (802.11b, a, and g) of the standard and thus is worth a moment's reflection. The MAC, although not as simple as the original Ethernet MAC, is nevertheless a fairly basic object. We have not discussed the point-coordination function because it is rarely used. Absent this, the MAC provides no central coordination of contending stations and no guarantees of performance except for the continuity of an ongoing packet exchange.

5.1.3.3 802.11 Classic direct-sequence PHY

To maximize the probability that multiple users could share the unlicensed bands without unduly interfering with each other, the FCC placed restrictions on communications systems that could operate there. A fundamental part of the requirement was that the systems not transmit all their energy in one narrow segment of the band but should spread their radiation over a significant part of the band, presumably a much larger segment than actually required by the data bandwidth, that is, users were required to apply *spread spectrum* techniques (a proviso since relaxed). In this fashion, it was hoped that interference between collocated systems would be minimized.

One approach to spreading the spectrum is to operate on many channels in some pseudo-random sequence; this is known as *frequency hopping* and is used in the classic FH PHY and the Bluetooth (802.15) PHY.[1] It is obvious that if two neighboring radios are operating on separate channels, interference ought to be minimized except when by chance they happen to select the same frequency. The disadvantage of this approach is that the channels must be quite narrow if there are to be a large number of them and therefore little chance of overlap; the 1-MHz maximum width specified in the regulations limits the data rate that can be transmitted at the modest signal-to-noise ratios (S/Ns) expected on low-cost, unlicensed, wireless links.

A quite distinct and rather more subtle approach is called *direct-sequence spread* DSSS. Direct-sequence methods were developed for military applications and are also used in those cellular telephones that are based on code-division multiple access (CDMA) standards. In DSSS, the relatively slow data bits (or more generally symbols) are multiplied by a much faster pseudo-random sequence of *chips*, and the product of the two is used as the transmitted signal. Recall that the bandwidth of a signal is determined by the number of symbols per second transmitted, whether or not those symbols contain useful information. Thus, the bandwidth of a DSSS signal is determined by the chip rate, which in general is much larger than the data rate; a DSSS signal can satisfy the FCC's requirement that the signal be spread. Further, the received signal can be multiplied again by the same sequence to recover the original lower data rate. In spectral terms, multiplying the received wide-bandwidth signal by the chip sequence collapses all its energy into a narrower bandwidth occupied by the actual data while simultaneously randomizing any narrowband interfering signal and spreading its energy out. Thus, the use of direct sequence codes provides *spreading gain*, an improvement in the link budget due to intelligent use of extra bandwidth beyond what is needed by the data.

In 802.11 classic DSSS, each data bit is multiplied by an 11-chip Barker sequence, shown in Figure 5.1-7. (Here data are shown as +1 or −1 both to assist the reader in thinking in terms of multiplication and due to the direct translation of the result into binary phase-shift keying [BPSK] transmitted chips; a formally equivalent treatment in terms of modulo-2 addition of binary bits is also possible.) Thus, a 1 bit becomes the sequence (+1−1 +1 +1 −1 +1 +1 +1 −1 −1 −1) and a 0 bit becomes (−1+1−1−1+1−1−1−1+1+1+1).

The Barker sequence is chosen for its *autocorrelation* properties: if two Barker sequences offset in time are

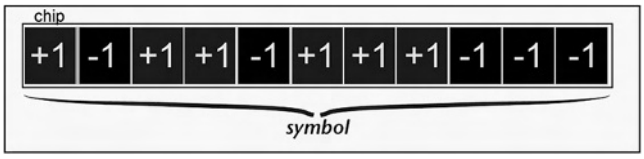


Figure 5.1-7 Barker sequence.

multiplied together and the result added (a *correlation* of the two sequences), the sum is small except when the offset is 0. Thus, by trying various offsets and checking the resulting sum, one can locate the beginning of an instance of the sequence readily; that is, synchronization to the transmitter is easy. It is also possible to have several DSSS signals share the same frequency band with little interference by using different *orthogonal* spreading codes for each signal: codes whose correlation with each other is small or zero. This technique is key to the operation of CDMA cellular phone standards but is not used in the 802.11 standards.

The basic symbol rate is 1 Mbps; each symbol consists of 11 chips, so the chip rate is 11 Mbps. The *spreading gain*, the ratio of the actual bandwidth to that required by the underlying data, is thus 11:1 or about 10.4 dB, meeting the FCC's original minimum requirement of 10 dB. Because the chip rate is 11 Mbps, one would expect the bandwidth required to be modestly in excess of 11 MHz. The actual bandwidth of the transmitted signal depends on the details of precisely how the signal is filtered and transmitted; rather than specifying such implementation-specific aspects, the standard simply defines a *spectral mask*, which provides a limit on how much power a compliant transmitter is allowed to radiate at a given distance from the nominal center frequency. The spectral mask for 802.11 classic is shown in Figure 5.1-8. The frequency reference is the nominal frequency for the given channel (about which we'll have more to say in a moment); the amplitude reference is the power density in the region around the nominal frequency. The boxes represent the maximum allowed power density; the smooth line is a cartoon of a typical power spectrum. Observe that although the standard allows the spectrum to be as much as 22 MHz wide, a typical real signal would have a bandwidth measured, for example, 10 dB from the peak of around 16 MHz.

The nominal center frequency for a transmission is chosen from a set of channel frequencies spaced by 5 MHz, shown in Table 5.1-1. (In the United States, channels 1–11 are available; channels 1–13 are allowed in most European jurisdictions, and channel 14 is for Japan.) Note that if the bandwidth is 16 MHz,

[1] In a remarkable example of unexpected ingenuity, frequency hopping was invented by the Austrian-born actress Hedy Lamarr, aided by composer George Antheil (U.S. patent 2,292,387), though she received little credit during her lifetime for this achievement.

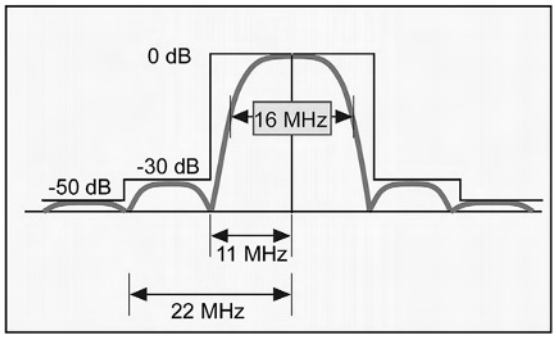


Figure 5.1-8 802.11 spectral mask.

transmissions on adjacent channels will interfere with each other quite noticeably. A separation of five channels (25 MHz) is needed to remove most overlap; thus, there are actually only three *nonoverlapping* channels available in the United States, channels 1, 6, and 11 (Figure 5.1-9).

Two data rates are supported in 802.11 classic: 1 and 2 Mbps. Both rates use the same chip rate of 11Mbps and have essentially the same bandwidth. The difference in data rate is achieved by using differing modulations. Packets at the basic rate of 1 Mbps, and the preambles of packets that otherwise use the extended rate of 2 Mbps, use differential BPSK (DBPSK), as shown in Figure 5.1-10. The data portion of extended-rate packets uses differential quaternary phase-shift keying (DQPSK). Recall that the bandwidth of a signal is mainly determined by the symbol rate, not by the nature of the individual symbols; because QPSK carries 2 bits per symbol, it is possible to roughly double the data rate without expanding the bandwidth, at a modest cost in required (S/N). In each case, the data bits are scrambled in order before transmission; this procedure avoids the transmission of long sequences of 1s or 0s that might be present in the source data, which would give rise to spectral artifacts.

The 802.11 modulations are differential variants of the BPSK and QPSK schemes, that is, the phase of each symbol is defined only with respect to the symbol that preceded it. Thus, if the phase of a BPSK signal is the same for two consecutive chips, a 0 has been transmitted.

Table 5.1-1 802.11 Channels.

Channel	*f* (GHz)
1	2.412
2	2.417
3	2.422
4	2.427
5	2.432
6	2.437
7	2.442
8	2.447
9	2.452
10	2.457
11	2.462
12	2.467
13	2.472
14	2.484

(Note that the arrows in Figure 5.1-10 showing the differences in phase are indicative only of the phase change between the moments at which the signal is sampled, not of the path the signal takes between sample times. This apparently subtle point has significant consequences in terms of the requirements on the radio. A moment's reflection will clarify why such a change might be desirable. Imagine that we wish to accurately state the absolute phase of a received QPSK signal by, for example, comparing it with a local oscillator that at the beginning of the packet is locked to the phase of the carrier. A 1500-byte (12,000 bit) packet at 2 Mbps lasts 6000 μsec. During this time, the total phase change of the carrier is 2π (2.4 GHz) $(6000 \times 10^{-6}\ \text{sec}) = 9 \times 10^{7}$ radians. If we wish the phase of the local oscillator to remain accurate to, for example, $\pi/12$ radians, so that this drift is small compared with the $\pi/2$ radian changes we're trying to

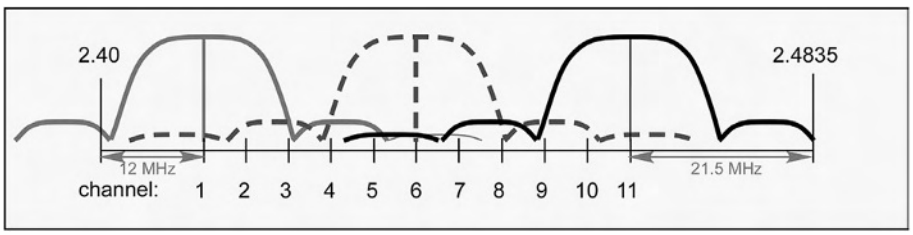


Figure 5.1-9 Nonoverlapping 802.11 channels, U.S. ISM band.

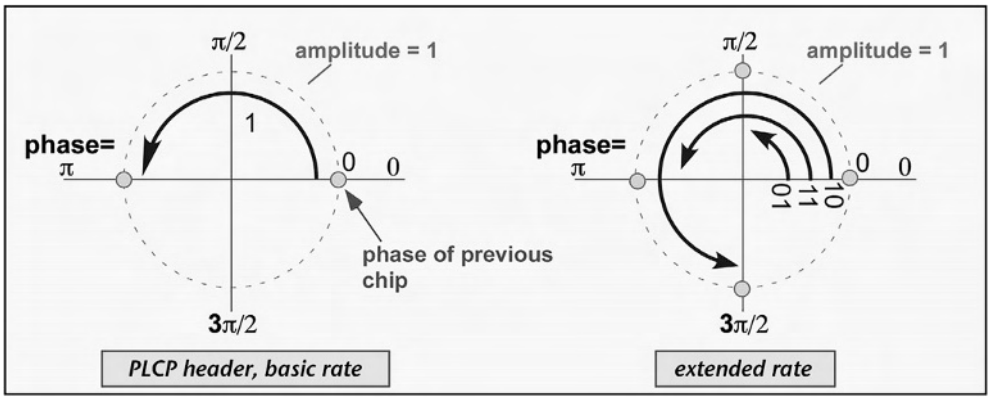


Figure 5.1-10 802.11 Classic modulations: DBPSK (left) and DQPSK (right).

resolve between chips, we need the oscillator to maintain phase to about 3 parts per billion. This is implausibly difficult and expensive. (Of course, more clever methods are available to perform such a task at more modest requirements on the hardware; the calculation is done to provide a frame of reference.) On the other hand, to maintain phase lock between successive chips at 11 Mbps, we only need hold this accuracy over 1/11 μsec: 220 cycles, or 1400 radians. The required accuracy in this case is 200 parts per million, which is easily achieved with inexpensive crystal-referenced synthesized sources. The price of this alternative is merely that we lose any information present in the first symbol of a sequence: because we can exploit our synchronization sequence (which doesn't carry any data anyway) for this purpose, it is obvious that differential modulation carries considerable benefits and little penalty.

The maximum transmit power of an 802.11 radio is limited in the standard to comply with the requirements of regulatory bodies. In the United States, power must be less than 1W, but in practice a typical access point uses a transmit power of about 30 to 100 mW. The standard then requires that a receiver achieve a frame error rate of less than 8% for 1024-byte frames of 2 Mbps QPSK at a signal power of –80 decibels from a milliwatt (dBm). That sounds like a lot of errors, but because each frame has $1024 \times 8 = 8192$ bytes, the bit error rate is a respectable $0.08/8192 = 10^{-5}$. To understand what the transmit and receive power levels imply, let's do a link budget calculation. Let the transmitted power be 100 mW (20 dBm). Assume the transmitting antenna concentrates its energy in, for example, the horizontal plane, achieving a 6-dB increase in signal power over that of an ideal isotropic antenna. Assume the receiving antenna has the same effective area as an ideal isotropic receiver. Our allowed path loss—the link budget—is then $(20 + 6 - (-80)) = 106$ dB. Recalling that an isotropic antenna has an equivalent area of around 12 cm^2, this is equivalent to a free space path of 2000 m or 2 km! However, recall that propagation indoors is unlikely to achieve the same performance as unimpeded ideal propagation in free space. We investigate the distinctions in more detail elsewhere; for the present let us simply add 30 dB of path loss for a long indoor path, leaving us with an allowed range of $(2000/10^{1.5}) = 63$ m. The transmit power levels and receiver sensitivities envisioned in the 802.11 standard make a lot of sense for indoor communications over ranges of 10–100 m, just as one might expect for a technology designed for the LAN environment (see section 5.1.1).

The receive sensitivity requirement also tells us something about the radio we need to build. Recall that a QPSK signal needs to have an (S/N) of about 12.5 dB for reasonably error-free reception. At first blush, one would believe this implies that the noise level of the receiver must be less than $(-80 - 12.5) = -92.5$ dBm. However, recall that 11 QPSK chips at 11 Mbps are "averaged" together (correlated with the Barker sequence) to arrive at one 2-bit data symbol. This averaging process provides us with some extra gain—the spreading gain described above—so that we can tolerate an (S/N) of $12.5 - 10.5 = 2$ dB and still get the data right. The noise level we can tolerate in the receiver is thus $-80 - 2 = -82$ dBm. Recalling that the signal is about 16 MHz wide, the unavoidable thermal noise in the receiver is $(-174\text{ dBm} + 10(\log 16) + 60) = -102$ dBm. The specification has left room for an additional 20 dB of excess noise in the receiver. Receivers with noise figures of 20 dB at 2.4 GHz are very easy to build, and in fact 802.11 commercial products do much better than this. The standard has specified a level of performance that can be reached inexpensively, appropriate to equipment meant for an LAN. The link budget calculation is displayed graphically in Figure 5.1-11.

5.1.3.4 802.11 Alphabet soup

Products implementing the 802.11 classic standard were manufactured and sold in the late 1990s by vendors such as Proxim, Breezecom (now part of Alvarion),

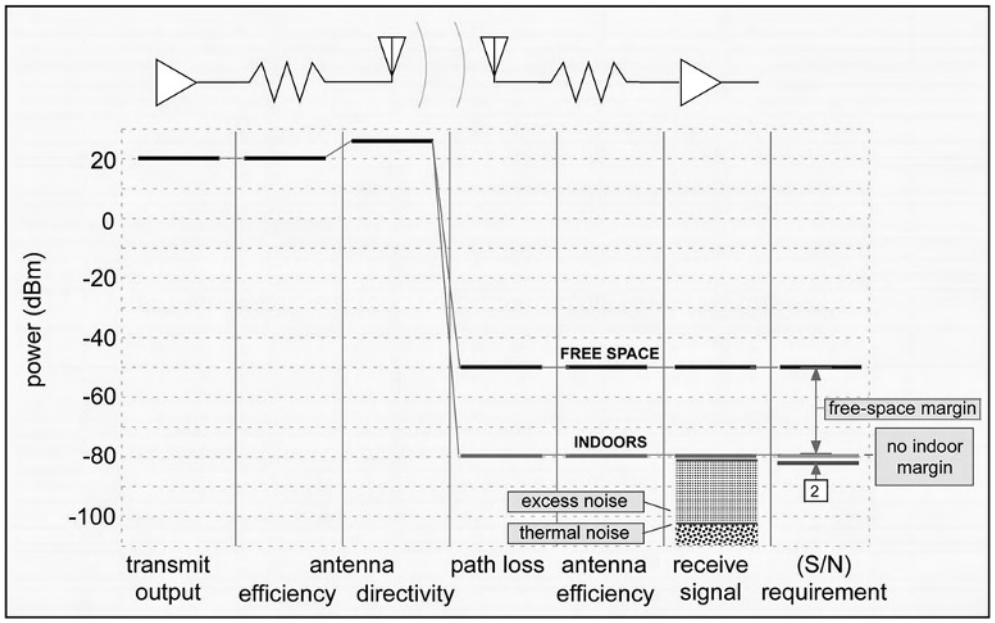


Figure 5.1-11 Link budget estimates for an 802.11–compliant radio link.

Lucent (under the WaveLAN brand), Raytheon, Symbol Technologies, and Aironet (now part of Cisco). However, the maximum data rate of 2Mbps represented a significant obstacle to the intended target usage model of wirelessly extending Ethernet: even the slowest variant of Ethernet has a native bit rate of 10 Mbps and a true throughput of around 8–9 Mbps. Work had started as early as 1993 on improved physical layers, and by 1999 two new higher performance PHY layers were released as enhancements to the original 802.11 classic interfaces. *802.11a* was a fairly radical modification of the physical layer, having essentially no commonalty with the DSSS or FH PHY layers of the classic standard and targeted for operation in a different slice of spectrum around 5 GHz. *802.11b* was a much more conservative attempt to improve performance of the 802.11 PHY layer without changing it very much. The 802.11b standard received the most early commercial attention, including support in the consumer market from Apple Computer's Airport product line and Cisco's Aironet products in the "enterprise" market (LANs for large industrial companies). The elaboration of compatibility efforts begun in association with the release of 802.11 classic led to the formation of the Wireless Ethernet Compatibility Alliance, or WECA, which (recognizing the awkwardness of the IEEE's nomenclature) provided *Wi-Fi* compliance certification to ensure interoperable 802.11b products from multiple vendors. By 2000 it was clear that 802.11b had found a "sweet spot," combining adequate data rates and performance with simplicity and low cost, and was poised for major commercial successes and widespread deployment. Sales volumes grew rapidly, and prices for both client cards and access points fell even more rapidly as numerous vendors entered the field.

The more audacious 802.11a standard, which would provide enhanced data rates of up to 54 Mbps, was seen by most industrial participants as a tool to hold in reserve for future use, but the start-up company Atheros proceeded to release chips implementing the 802.11a standard, with products available from Intel and Proxim in 2002. These products, although modestly successful at introduction, clarified the desirability of combining the high data rates of the 802.11a standard with backward compatibility to the installed base of 802.11b products. The "G" task group, which was charged to take on this awkward task, after some struggles and dissension approved the 802.11g standard in 2003, noticeably after the release of several prestandard commercial products. At the time of this writing (early 2004), 802.11g products have nearly replaced 802.11b products on the shelves of consumer electronics stores: the promise of backward compatibility in conjunction with a much higher data rate at nearly the same price is an effective sales incentive, even though many consumer applications do not currently require the high data rates provided by the standard. Dual-band access points and clients, supporting both 802.11a and 802.11b/g, are also widely available.

A number of task groups are also actively attempting to fill in holes in the original standards work and add new capabilities. Task group "I" is charged with improving the limited authentication and encryption capabilities provided in the original standard, about which we have more to say in section 5.1.3.8. Task group "F" is responsible for providing the missing description of the distribution

system functions, so that roaming between access points can be supported across any compliant vendor's products; an IAPP specification was approved in June 2003.

Recall that the original Ethernet standard, and the 802.11 MAC derived from it, deliver *best-effort* data services, with no guarantees of how long the delivery will take. Services such as voice or real-time video delivery require not only that data be delivered but that it arrive at the destination within a specified time window. This requirement is commonly placed under the not-very-informative moniker of *quality of service*. Task group "E" is defining quality of service standards for time-sensitive traffic over 802.11 networks, although to some extent that effort has refocused on the 802.15 work we describe in the next section. Task group "H" is defining two important additions to the 802.11a standard: *dynamic frequency selection* and *transmit power control*. These features are required for operation in most European jurisdictions, and the recent FCC decisions on spectrum in the United States will also require such capability for operation in most of the 5-GHz bands.

In the next three subsections we examine the important 802.11b and 802.11a PHY layers and briefly touch on their admixture in 802.11g. We also provide a cursory discussion of WLAN (in)security and some cures for the deficiencies revealed therein. We must regrettably refer the reader to the IEEE web sites described in section 5.1.3.7 for more details on the other task groups within 802.11.

5.1.3.5 The Wi-Fi PHY (802.11b)

The 802.11b physical layer uses the same 2.4-GHz band and channelization as the classic PHY. Furthermore, the basic signaling structure of 11 megasamples/second (Msps) of either BPSK or QPSK symbols is unchanged, and so it may be expected that the frequency spectrum of the transmitted signals will be similar to those of the classic PHY. However, the use of these symbols is significantly different.

To maintain compatibility with classic systems, packets with preambles transmitted at the lowest rate of 1 Mbps, using DBPSK modulation, must be supported. However, the new PHY adds the option to use short preambles with 2 Mbps DQPSK modulation to reduce the overhead imposed by the very slow long preamble on short high-rate packets. More importantly, the 802.11b PHY introduces two completely new approaches to encoding the incoming data onto the QPSK symbols: *complementary code keying* (CCK) and *packet binary convolutional coding* (PBCC). Each method may be used to support two new data rates, 5.5 and 11 Mbps. Both methods completely abandon the Barker sequence and conventional direct-sequence spreading.

Let us first examine CCK. CCK is a *block code:* chunks of symbols of fixed size are used as code words, and the subset of *allowed* code words is much smaller than the total *possible* set of code words. Errors are detected and corrected by comparing the received code word with the possible code words: if the received code word is not an allowed word but is close to an allowed word, one may with good confidence assume that the nearby allowed code word is in fact what was transmitted. Block codes are relatively easy to implement.

We look at the high-rate 11 Mbps coding in detail. In the particular code used in the 802.11b CCK-11 Mbps option, transmitted QPSK symbols are grouped into blocks of eight to form code words. Because each symbol carries 2 bits, there are $4^8 = 65{,}536$ possible code words. Of this large domain, only 256 code words, corresponding to the 8 input bits that define the chosen CCK block are allowed. They are found in the following fashion, shown schematically in Figure 5.1-12. (Figure 5.1-12 is headed "even-numbered symbol" because alternate code words have slightly different phase definitions, important for controlling the spectrum of the output but confusing when one is trying to figure out how the scheme works.) The input bits are grouped in pairs (dibits). Each dibit defines one of four possible values of an intermediate phase φ. The intermediate phases are then added up (modulo-2π) as shown in the chart to determine each QPSK symbol c of the transmitted code word. Thus, the last symbol c7 shares the phase of $\varphi 1$ (and is used as the phase reference for the remainder of the code word).

The phase of c4 is $(\varphi 1 + \varphi 2 + \varphi 3)$. Symbols c3 and c6 have π radians added to the argument (i.e., they are multiplied by $\varphi 1$).

This procedure is a bit confusing to describe but easy to implement and ensures that the resulting allowed code words are uniformly distributed among all the possible code words. Recall that we defined the distance between two QPSK symbols by reference to the phase-amplitude plane; for a normalized amplitude of 1, two nearest neighbors are $2/\sqrt{2}$ apart; if we square the distance we find nearest neighbors differ by 2 in units of the *symbol energy*, E_s. We can similarly define the difference between code words by adding up the differences between the individual QPSK symbols, after squaring to make the terms positive definite (> 0). Having done so, we find that CCK-11 code words have their nearest neighbors at a squared distance of 8. This means that no possible single-chip QPSK error could turn one allowed code word into another, so single-chip errors can always be detected and corrected. In RF terms, the bit error rate is reduced for a fixed (S/N), or equivalently the (S/N) can be increased for the same error rate. The change in (S/N) is known as *coding gain*. The coding gain of CCK is about

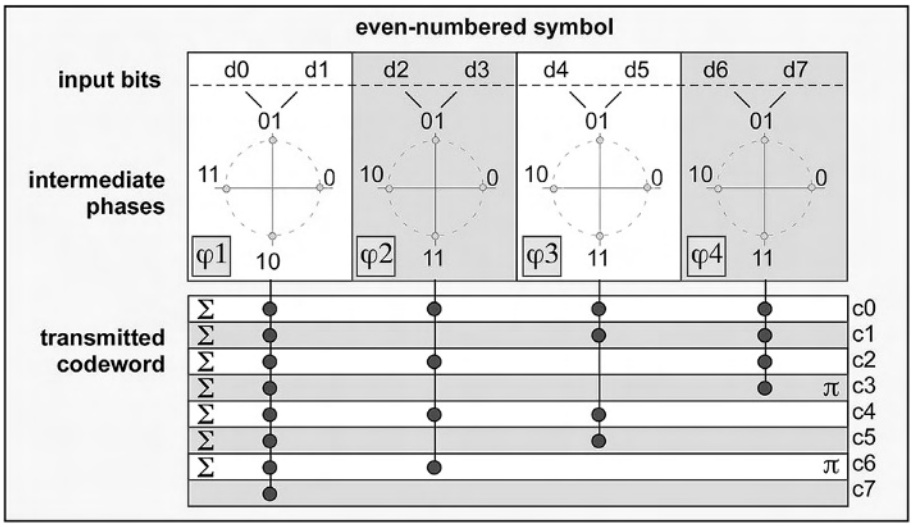


Figure 5.1-12 CCK-11 code word encoding scheme (even-numbered symbols).

2 dB. A slight variant of the above scheme is used to deliver 5.5 Mbps.

The standard also provides for a separate and distinct method of achieving the same 5.5- and 11-Mbps rates: PBCC. PBCC is based on a *convolutional code* (Figure 5.1-13). The coding is performed using a series of shift registers, shown here as z^1 through z^6. At each step, a new input bit is presented to the coder and all bits are shifted to the right one step. The "+" symbols represent modulo-2 additions that produce the output bits y_0 and y_1 from the input and the stored bits. The coder can be regarded as a machine with $2^6 = 64$ possible states. In any given state of the coder, two transitions to the next state are possible (which occurs being determined by the new input bit) and two of the four possible outputs are allowed (again determined by the input). Thus, the sequence of allowed outputs is highly constrained relative to the sequence of possible outputs. This code is known as a *rate 1/2* code, because each input bit produces two output bits.

The output bits are mapped onto the next QPSK symbol in two possible ways (rotated 90 degrees with respect to each other) according to a 256-bit fixed sequence. Because there are 11 megachips per second and each chip carries one input data bit coded into two y's, the net data rate is 11 Mbps. The 256-bit sequence helps to remove any periodicity in the output signal and thus helps smooth the spectrum of the output. The relatively long period of the sequence means that any other interfering transmission is unlikely to be synchronized with it and will appear as noise rather than as valid code words, making it easier to reject the interference.

Convolutional codes are usually decoded with the aid of a *Viterbi trellis* decoder. (Trellis algorithms can also be used to decode block codes.) The trellis tracks all the possible state transitions of the code and chooses the trajectory with the lowest total error measure. It would seem at first glance that such a procedure would lead to an exponentially growing mess as the number of possible trajectories doubles with each additional bit received,

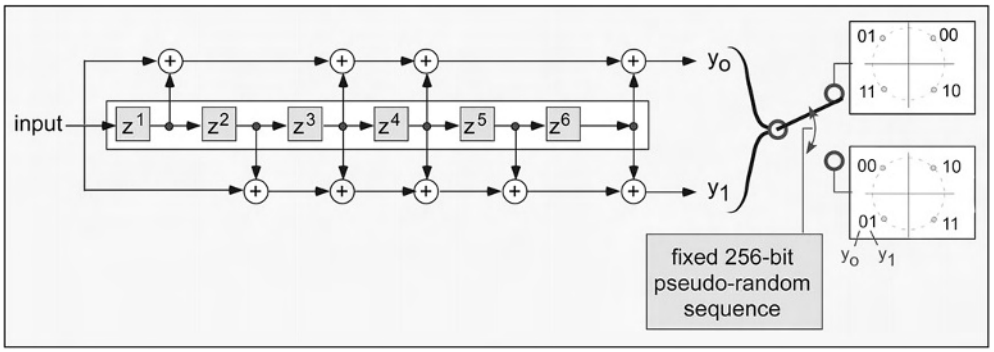


Figure 5.1-13 PBCC-11 convolutional code and cover sequence.

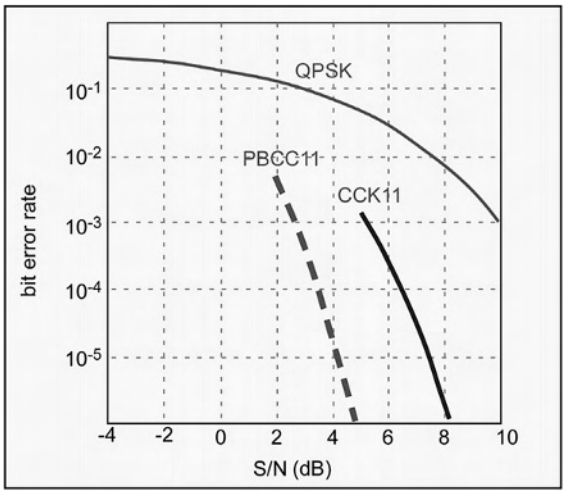


Figure 5.1-14 Error rate of uncoded QPSK, PBCC11, and CCK 11 vs. (S/N) (after Heegard et al., in Bing 2002).

but by pruning out the worst choices at each stage of the trellis, the complexity of the algorithm is reduced to something manageable. However, implementation is still rather more complex than in the case of a block code like CCK.

The PBCC-11 code has slightly better performance than the corresponding CCK code: about 3.5 additional dB of coding gain (Figure 5.1-14). However, the computational complexity of decoding the convolutional code is about 3.5 times larger than the CCK code. To the author's knowledge, the PBCC variant has been implemented only in a few products from D-Link and US Robotics and has enjoyed very little commercial success. Commercial politics may have played a significant role in this history, as Texas Instruments enjoyed certain intellectual-property rights in PBCC that may have made other vendors reluctant to adopt it. It is also important to note that in most applications of a WLAN, performance is secondary to price and convenience; the modest advantage in link budget of PBCC was not enough to produce a commercial advantage in the marketplace.

5.1.3.6 802.11a PHY

The 802.11a PHY is a radical departure from the approaches above. The first major change is the use of the Unlicensed National Information Infrastructure (UNII) band at 5.15–5.825 GHz instead of the 2.4-GHz ISM band. At the time the PHY was promulgated, 5-GHz radios implemented in standard silicon processing were not widely available, so this choice implied relatively high costs for hardware. The motivation for the change was the realization that a successful commercial implementation of 802.11 classic and/or 802.11b devices, combined with the other proposed and existing occupants of the 2.4-GHz band (Bluetooth devices, cordless telephones, and microwave ovens, among others), would eventually result in serious interference problems in this band. The UNII band was (and is) relatively unoccupied and provides a considerable expansion in available bandwidth: in the United States, 300 MHz at the time versus 80 MHz available at ISM.

Rules for operating in the UNII band have changed significantly in the United States since the promulgation of the 802.11a standard: in November 2003, the FCC added 255 MHz to the available spectrum and imposed additional restrictions on operation in the existing 5.250- to 5.350-GHz band. The original and modified U.S. assignments are depicted in Figure 5.1-15. The allowed bands at the time of the standard consisted of a lower, middle, and upper band, each with 100 MHz of bandwidth. The lower band was dedicated to indoor use only and limited to 40 mW output power. The middle band allowed dual use, and the upper band was targeted to outdoor uses with a much higher allowed output of 800 mW.

The FCC's changes in late 2003 added an additional band, almost as big as what had been available, for dual use at 200 mW. Use of this band, and retroactively of the old UNII mid-band, is only allowed if devices implement transmit power control and dynamic frequency selection. The former reduces the power of each transmitting device to the minimum needed to achieve a reliable link, reducing overall interference from a community of users. The latter causes devices to change their operating frequency to an unoccupied channel when possible. These changes in band definitions and usage requirements improve consistency between U.S. and European (ETSI) requirements in the 5-GHz band and will presumably increase the market for compliant devices. The 802.11h standard, approved in 2003 to allow compliant devices that meet European standards, adds several capabilities relevant to power control and channel management. Clients inform access points about their power and channel capabilities; quiet periods are added when no station transmits to enable monitoring of the channel for interference, and access points can use the point coordination function interframe space to capture the medium to coordinate switching channels when interference is detected.

Devices that are compliant to 802.11h should also meet the new FCC standards for UNII operation in the United States.

The spectral mask limiting the bandwidth of the transmitted signal is shown in Figure 5.1-16. Note that although the mask is somewhat more complex and apparently narrower than the old 802.11 mask (Figure 5.1-8), in practice both signals end up being around 16-MHz wide at 10 dB down from the maximum intensity at the center of the transmitted spectrum. This design choice was made to allow the same analog-to-digital conversion hardware to be used for the baseband in either standard, which is beneficial in those cases where a single baseband/

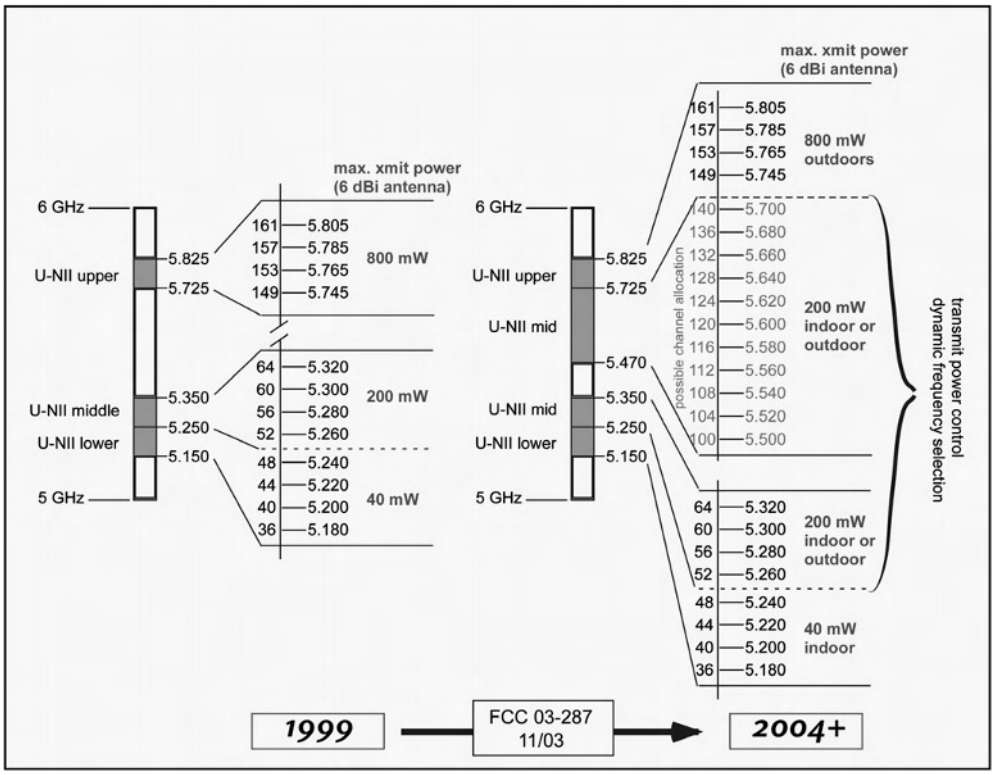


Figure 5.1-15 UNII band in the United States, then and now.

MAC design is to serve both 802.11a and 802.11b radio chips. This important fact allowed the 802.11a PHY to be kidnapped and transported whole into the ISM band in the 802.11g standard, as we discuss in section 5.1.3.7.

Nonoverlapping channels in the old band definition, and a possible channel arrangement in the new bands, are shown in Figure 5.1-17. The assumed channel shape is based on the spectral mask in Figure 5.1-16. We can see that the old definition had room for as many as 12 nonoverlapping channels, though operation at the high band at full power in the lowest and highest channels would result in out-of-band emissions exceeding FCC allowances, so that in practice fewer channels might be used. In the new band assignment, there is room for up to 19 nonoverlapping channels in the lower and middle bands alone. Recall that the ISM band allows only three nonoverlapping 802.11 channels (Figure 5.1-9).

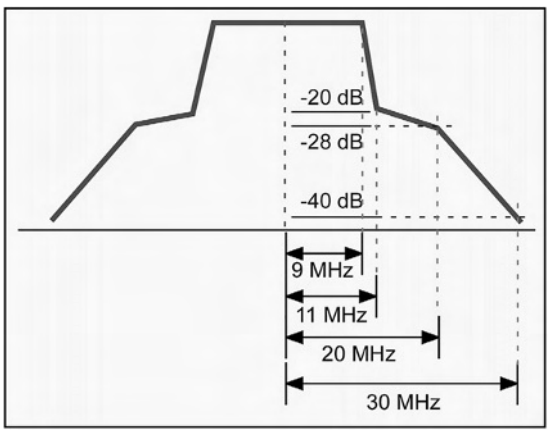


Figure 5.1-16 802.11a Transmit spectral mask.

The availability of more than three channels is a significant benefit for implementing networks that are intended to provide complete coverage to a large contiguous area. With four nonoverlapping channels, one can construct a network of individual access points according to *a frequency plan* that ensures that access points on the same channel are separated by at least four cell radii; with seven channels to work with, a minimum spacing of 5.5 radii is achieved (Figure 5.1-18). Thus, interference between adjacent cells is minimized, and each cell is able to provide full capacity to its occupants. With the large number of independent channels now available, individual access points can be allocated more than one channel each to increase capacity in heavily used cells while still maintaining minimal interference.

Coding in 802.11a is somewhat similar to the PBCC option of the 802.11b PHY: after scrambling, incoming bits are processed by rate 1/2 convolutional encoder as shown in Figure 5.1-19. Like the PBCC encoder, the 802.11a code has a six-stage shift register, though the sequence of taps that define the output bits (formally, the *generating polynomials* of the code) have been changed. Codes with rates of 3/4 and 2/3 (fewer output

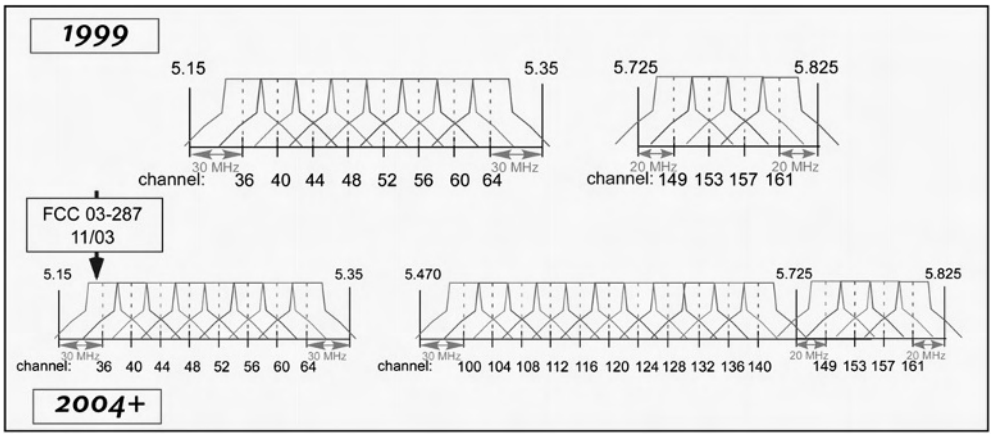


Figure 5.1-17 Possible nonoverlapping channel assignments in old and new UNII bands.

bits per input bit and thus higher throughput of useful data) are created by *puncturing* the output of the rate 1/2 code, that is, some of the output bits are simply ignored. Such a procedure is roughly equivalent to intentionally introducing some bit errors in the data stream, though it is easier to correct for because the location of the missing bits is known; if done cleverly, it allows for higher throughput with little performance loss.

The use of the resulting output bits is significantly different from PBCC, because instead of single-carrier modulation, the 802.11a standard uses orthogonal frequency-division multiplexing (OFDM). A data symbol is composed of 48 subcarriers, which carry data, and 4 additional *pilot* subcarriers, which transmit a known pseudo-random sequence to assist the receiver in maintaining synchronization with the transmitter. The subcarriers are assigned to 64 frequency slots, separated by 312.5 KHz; the slots at the band edges and the slot in the center of the band (i.e., at the carrier frequency) are not used. If all the subcarriers were occupied, the signal would be approximately 64×312.5 kHz = 20 MHz wide (63 spacings between subcarriers + half of the width of a subcarrier on each side); the actual (ideal, undistorted) signal is about 16.6 MHz wide, thus fitting nicely within the spectral mask in Figure 5.1-16.

Recall that to use OFDM, the time over which a symbol is integrated must be an integer number of cycles for all the subcarriers in order that orthogonality is maintained. The native symbol time of the 802.11a OFDM symbol is thus one cycle of the lowest frequency subcarrier, or 3.2 μsec. A cyclic prefix is appended to the symbol to allow for a guard interval that eliminates the effects of multipath. The cyclic prefix is 0.8 μsec long (about 20% of the total symbol), so that the total symbol length is 4 μsec. Therefore, the OFDM symbol rate is 250 Ksps. Each symbol contains 48 active subcarriers, so this is equivalent to sending 12 million single-carrier symbols per second.

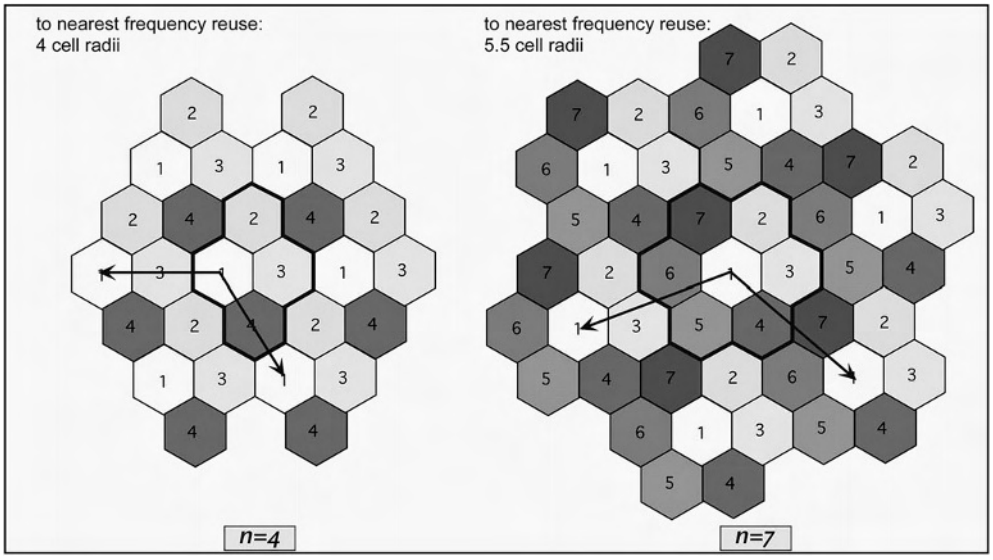


Figure 5.1-18 Examples of cellular coverage plans for 4 and 7 nonoverlapping channels.

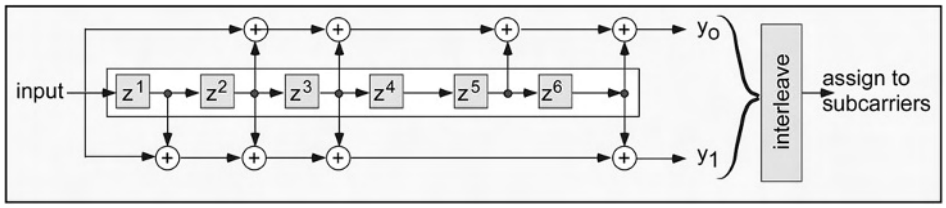


Figure 5.1-19 802.11a Convolutional encoder.

To allow for varying conditions, many different combinations of code rate and subcarrier modulation are allowed. These are shown in Table 5.1-2. Recall that BPSK transports one bit per (subcarrier) symbol, QPSK two bits, and so on up to 64 quadrature-amplitude-modulation (QAM), which carriers 6 bits per symbol. The product of the bits per symbol of the modulation and the rate of the convolutional code gives the number of bits each subcarrier contributes to each OFDM symbol; multiplication by 48 provides the number of bits transported per OFDM symbol.

Note that to avoid bursts of bit error from disruption of a few neighboring subcarriers, the output bits from the code are *interleaved*—that is, distributed in a pseudo-random fashion over the various subcarriers.

Several views of a simulated OFDM frame are shown in Figure 5.1-20. The frame consists of a preamble and simulated (random) data. The spectrum of the output signal is about 16-MHz wide, in good agreement with the simple estimate given above. Recall that a signal in time can be described by its in-phase and quadrature components, I and Q. The I and Q amplitudes for the frame are shown at the lower left. The frame begins with a set of simplified synchronization characters, which use only 12 subcarriers, and are readily visible at the left of the image. The remainder of the preamble and data frame appear to vary wildly in amplitude. This apparently random variation in the signal is also shown in the close-up of I and Q amplitudes over the time corresponding to a single OFDM symbol shown at the top right of the figure.

Table 5.1-2 802.11a Modulations and code rates.

Modulation	Code rate	Data rate (Mbps)
BPSK	1/2	6
BPSK	3/4	9
QPSK	1/2	12
QPSK	3/4	18
16QAM	1/2	24
16QAM	3/4	36
64QAM	2/3	48
64QAM	3/4	54

At the bottom right we show the *combined cumulative distribution function* for the signal power, averaged over the packet length, for a packet at 54 and 6 Mbps. The combined cumulative distribution function is a graphic display of the frequency with which the instantaneous power in the signal exceeds the average by a given value. The 54-Mbps packet instantaneous power is greater than 9 dB above the average around 0.1% of the time. The 6-Mbps packet displays an 8-dB enhancement in power with this frequency of occurrence. Although that may not seem like much, recall that the spectral mask (Figure 5.1-16) requires that the transmitted signal 20 MHz away from the carrier must be reduced by more than 28 dB. A high-power signal may become distorted, resulting in spectral components far from the carrier. If such components were comparable in size with the main signal and occurred only 1% of the time, their intensity would be 20 dB below the main signal. Infrequent power peaks contribute significantly to the output spectrum when distortion is present. This problem of a high ratio of peak power to average power represents one of the important disadvantages of using OFDM modulations, because it forces transmitters to reduce their average output power and requires higher precision and linearity from receivers than is the case for simpler modulations.

Recall that the OFDM symbol is constructed using an inverse fast Fourier transform (FFT), with the reverse operation being performed on the received signal to recover the subcarriers. In the case of 802.11a, a 64-entry FFT involving some hundreds of arithmetic operations must be performed at fairly high precision every 4 μsec, followed by a trellis decode of the resulting 48 data symbols. The rapid progress of digital integrated circuit scaling has permitted such a relatively complex computational system to be implemented inexpensively in standard silicon complementary metal-oxide semiconductor (CMOS) circuitry, a remarkable achievement.

5.1.3.7 802.11g PHY

The great disadvantage of the 802.11a PHY is its incompatibility with the older classic and 802.11b installed base. To overcome this obstacle while preserving the

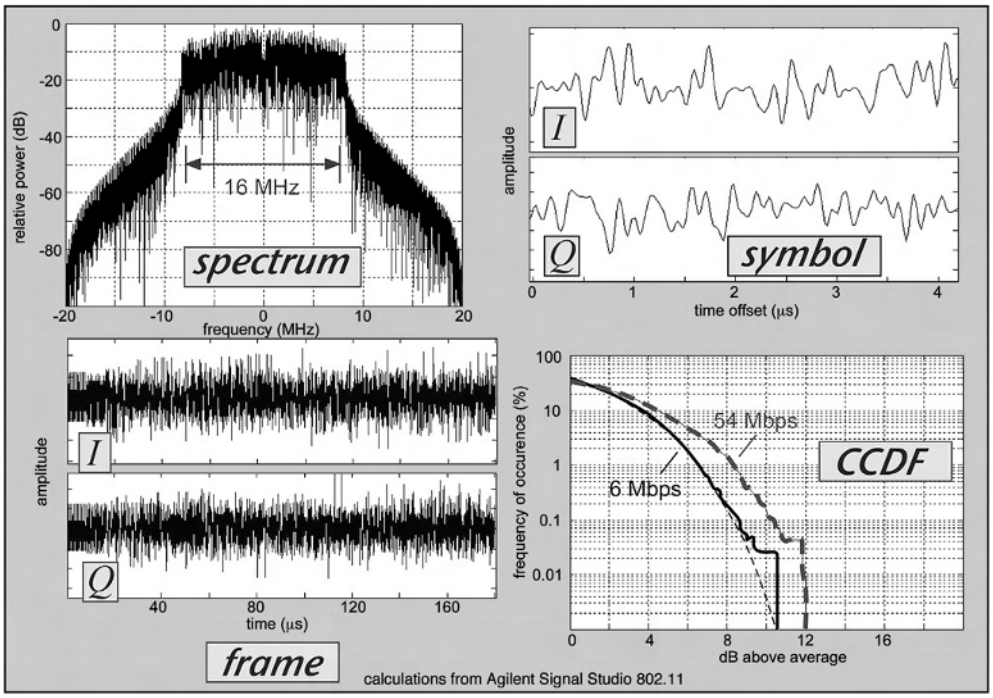


Figure 5.1-20 Several views of an 802.11 OFDM frame.

attractive high peak data rates of the 802.11a standard, task group "G" provided two enhancements to the Wi-Fi PHY. The first variant is essentially the transfer of the 802.11a OFDM physical layer *in toto* to the ISM band. This trick is possible because the bandwidth of the 802.11a symbols (Figure 5.1-20) is about 16 MHz, just as the classic symbols were.

The problem presented by this radical reinvention is that older systems have no ability to receive and interpret the complex OFDM symbols and simply see them as noise. Because the CSMA/CA MAC layer is heavily dependent on all stations being able to hear (at least) the access point, invisible preambles are a serious obstacle: receiving stations cannot set their NAV to reserve the medium (Figure 5.1-6) if they can't read the packet preambles. In order for "g" stations to coexist with older "b" stations, several options are possible. First, in the presence of mixed traffic, "g" stations can use the RTS/CTS approach to reserve the medium using 802.11b packets and then use the reserved time to send faster 802.11g packets. This approach is simple but adds overhead to the "g" packet exchanges. Alternately, a station can send a CTS packet with itself as the destination: this approach, known as a CTS-to-self, is obviously more appropriate for an access point, which expects that every associated station can hear it, than a client.

A "g" station can also use a mixed frame format, in which the preamble is sent in a conventional 802.11b form and then the appended data are transferred as OFDM. In this case there is additional overhead because of the low data rate of the older preamble but no additional packet exchange. The final chip of the preamble is used as the phase reference for the first OFDM symbol.

Finally, "g" also defines a PHY using an extension of the PBCC coding system; in the highest-rate variant, 8PSK is substituted for QPSK modulation. It is perhaps too early in the history of 802. 11g to draw confident conclusions about its evolution, but in view of the limited success of PBCC in 802.11b deployment, and successful if limited deployment of OFDM-based 802.11a chipsets by major manufacturers, it seems plausible that the OFDM variant will dominate 802.11g devices.

Trade-offs are necessarily encountered in the enhancement of the PHY. Higher data rates use higher modulation states of the subcarriers and thus require better (S/N) than the old QPSK chips. This implies that higher rate communications will have a shorter range than the lower rates. The use of OFDM signals gives rise to higher peak-to-average power ratios, allowing less transmission power from the same amplifiers to avoid distortion. The benefit of coexistence is achieved at the cost of reduced actual data rates in the presence of legacy stations due to the overhead of any of the coexistence protocols discussed above. These problems are fairly minor bumps in the road, however: it seems likely that 802.11g products will dominate the WLAN marketplace until crowding in the ISM band forces wider adoption of 802.11a/dual-band devices.

5.1.3.8 802.11 (In)Security

Most participants in wireless networking, or indeed networking of any kind, are at least vaguely aware that 802.11 presents security problems. One can hardly have a discussion of 802.11 without dealing with encryption and security; however, as whole books have already been written on this subject, the treatment here will be quite cursory.

We must first put the problem in its wider context. Most enterprise LANs have minimal internal authentication and no security, because it is presumed that physical security of the Ethernet ports substitutes for systemic provisions of this nature (though those hardy souls charged with administering networks at universities may find such presumptions laughable at best). Access to the network from the outside, however, is typically limited by a firewall that controls the type and destination of packets allowed to enter and exit the local network. Remote users who have a need to use resources from an enterprise network may use a *virtual private network* (VPN), which is simply a logical link formed between two clients, one at the remote site and the other at the enterprise network, with all traffic between them being encrypted so as to defeat all but the most determined eavesdropper. Traffic on the Internet is similarly completely open to interception at every intermediate router; however, reasonably secure communications over the Internet can be managed with the aid of the Secure Sockets Layer (SSL), another encrypted link between the user of a web site and the site server.

It was apparent to the developers of wireless data devices that the wireless medium is more subject to interception and eavesdropping than the wired medium, though they were also aware of the many vulnerabilities in existing wired networks. They therefore attempted to produce a security mechanism that would emulate the modest privacy level obtained with physically controlled Ethernet wired ports in a fashion that was simple to implement and would not incur so much overhead as to reduce the WLAN data rate by an objectionable amount.

The resulting *Wired Equivalent Privacy* (WEP) security system was made optional in the classic standard but was widely implemented and extended. The system is a *symmetric key* system, in which the same key is used to encrypt and decrypt the data, in contrast to asymmetric systems in which a public key is used to encrypt data and a separate private key is used to decrypt it. Public key systems are much more difficult to penetrate but very computation-intensive and are used for the secure exchange of symmetric keys but not for packet-by-packet encryption. The basic idea of the encryption approach is to use the key to create a pseudo-random binary string (the *cipher stream*) of the same length as the message to be encrypted. The data bits (the *plaintext*) can then be added modulo-2 (which is the same as a bit-by-bit exclusive or [XOR]) to the cipher stream to create the encrypted data or *cipher text*, which when added again to the same cipher stream will recover the original data (plaintext). The general scheme is shown in Figure 5.1-21.

A simple way to achieve such a result is to use the key as the seed value for a random number generator algorithm. The problem with this naive implementation is that every instance of the (pseudo)random sequence would then be identical; it wouldn't be very hard for an eavesdropper to reconstruct the cipher stream once a large number of packets, many containing known plaintext such as the contents of standard Internet packets, had been intercepted. Therefore, to avoid using the same cipher stream for every packet while also avoiding the complexity of exchanging a new key between the sender and receiver each time a packet needs to be transmitted, the WEP algorithm uses an *initialization vector* (IV): 24 additional bits that are prepended to the 48-bit secret key to form the seed for the random number generator. In this fashion, each new packet gets

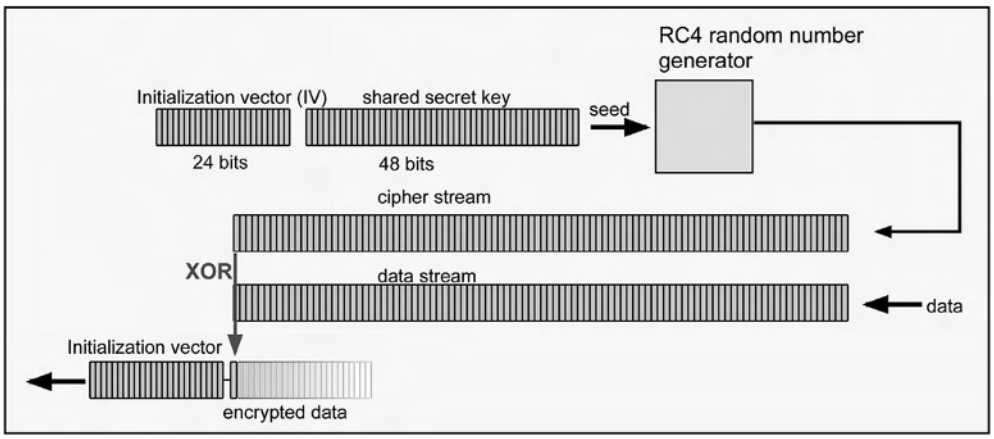


Figure 5.1-21 Simplified schematic of WEP encryption.

a new seed for the random number generator and thus a unique cipher stream.

In addition to security from interception, a secure communications system should provide some means of *authentication:* each side should have some confidence that they are in fact communicating with the machine (and ultimately the person) that is intended. The 802.11 standard appropriates WEP to support authentication by allowing an access point to challenge a potential client with a plaintext packet, which the client then returns encrypted with the shared WEP key. If the decrypted packet matches the original plaintext, the client is presumed to have knowledge of the shared key and thus be privileged to make use of the resources of the access point.

As has been so vividly shown with the public release of the remarkable means by which British and American researchers cracked the German Enigma code during World War II, attacks on cryptographic systems are rarely frontal but instead exploit weaknesses in the system and implementation. WEP has a number of weaknesses and limitations, some of which were quickly apparent and others more subtle and only revealed after significant work by outside researchers.

The first weakness is in the definition of the IV. The IV is necessarily sent in the clear because it is needed to decrypt the message content. As we noted above, to get decent security out of a random-number-generator scheme, it is necessary to avoid reusing the seed value, because this will generate a reused cipher stream. However, the IV space is only 24 bits. Furthermore, there is no requirement in the specification about how the IV is to be treated between successive packets; any IV value may be used and must be accepted by the receiving station. Many implementations simply incremented the IV value by 1 for each packet sent. In such a network, the IV is bound to be reused after $2^{24} = 16.8$ million packets. A heavily loaded network would deliver that much traffic in a couple of hours. Over the course of a week or two of eavesdropping, a huge number of packets with identical IVs and thus identical cipher streams could be collected and analyzed.

Once any cipher stream is obtained, it can be reused even in the absence of knowledge of the shared key, because any IV must be accepted. Any packets sent with the same IV can be decrypted (at least to the length of the known cipher stream); further, a packet of the same length with the intruder's data can be encrypted, prepended with the same IV, and injected into the system.

The authentication system is another weakness. If an eavesdropper can hear the challenge and the response, the XOR of the challenge (plaintext) and corresponding response (cipher text) reveals the cipher stream. In this fashion one can accumulate cipher streams for packet injection or other attacks.

Obviously, it is even better to obtain the shared key and thus encrypt and decrypt freely than to extract a few cipher streams. Attacks to find the shared key can exploit the fact that the key is defined as a bit stream, but human users can't remember bit streams very well and tend to prefer words or streams of recognizable characters. Many 802.11 management systems allow the entry of a text password, which is processed (hashed) into a 48-bit shared key. Early hashing algorithms did not use the whole key space—that is, not all 48-bit numbers were possible results—so that the effective size of the keys to be searched was only 21 bits, which is about 2 million possibilities. (Some early cards even converted the ASCII values of letters directly into a key, resulting in a greatly reduced key space that is also easily guessed.) A very simple brute force recitation of all possible passwords is then possible, given that one can automate the recognition of a successful decryption. Because certain types of packets (standard handshakes for IPs) are present in almost any data stream, the latter is straightforward.

Because IVs can be reused, if there isn't enough traffic on a network to crack it, an attacker can always introduce more packets once he or she has acquired a few cipher streams, thus generating additional traffic in response.

Certain attacks are fixed or rendered much less effective by simply increasing the length of the shared key and the IV. Many vendors have implemented WEP variants with longer keys, though these are not specified in the standard and thus interoperability is questionable. For example, the larger space for keys, combined with improved hashing algorithms, renders the brute force password guessing attack ineffective on 128-bit WEP key systems.

The widely publicized "cracking" of the WEP algorithm by Fluhrer, Mantin, and Shamir (FMS) was based on a demonstration that the particular random number generator, RC4, used in WEP is not completely random. RC4 is a proprietary algorithm that is not publicly released or described in detail in the standard. It has been reverse engineered and shown to be based on fairly simple swaps of small segments of memory. Fluhrer and colleagues showed that certain IV values "leak" some information about the key in their cipher streams. To benefit from this knowledge, an attacker needs to know a few bytes of the plaintext, but because of the fixed nature of formats for many common packets, this is not difficult. The number of weak IVs needed to extract the shared key is modest, and the frequency of weak IVs increases for longer key streams, so that for a network selecting its IVs at random, the time required to crack, for example, a 104-bit key is only roughly twice that needed for a 40-bit key. Long keys provide essentially no added security against the FMS attack. Shortly after Fluhrer et al. became available, Stubblefield, Ioannidis, and Rubin implemented the algorithm and showed that keys could be extracted after interception of on the order of 1,000,000 packets.

Finally, we should note that there is no provision whatsoever in the standard to allow a client to authenticate the access point to which it is associating, so that an attacker with a shared key can set up a spoof access point and intercept traffic intended for a legitimate network.

All the above attacks depend on another weakness in the standard: no provision was made for how the shared secret keys ought to be exchanged. In practice, this meant that in most systems the secret keys are manually entered by the user. Needless to say, users are not eager to change their keys every week, to say nothing of every day, and manual reentry of shared keys more than once an hour is impractical. Further, manual key exchange over a large community of users is an administrative nightmare. As a consequence, it is all too likely that WEP shared secret keys will remain fixed for weeks or months at a time, making the aforesaid attacks relatively easy to carry out.

Some of the problems noted above can be solved completely within the existing standards. Weak IVs can be avoided proactively, and some vendors have already implemented filters that ensure that weak IVs are never used by their stations. If all stations on a network avoid weak IVs, the FMS attack cannot be carried out. Weak hashing algorithms for short keys can be avoided by entering keys directly as hexadecimal numbers. In home networks with few users and relatively light traffic, manual key exchange on a weekly or even monthly basis will most likely raise the effort involved in a successful attack above any gains that can be realized.

The simple measures cited above can hardly be regarded as adequate for sensitive industrial or governmental data. In 2003, products conforming to a preliminary standard promulgated by the WECA industrial consortium, known as Wi-Fi Protected Access (WPA), became available. WPA is a partially backward-compatible enhancement of WEP. WPA uses the IEEE 802.1x standard as a framework for authentication of users and access points; within 802.1x various authentication algorithms with varying complexity and security can be used.

WPA uses a variant of the *Temporal Key Integrity Protocol* created by Cisco to improve security of the packet encryption process. An initial 128-bit encryption key, provided in a presumably secure fashion using the 802.1x authentication process, is XOR'd with the sending station's MAC address to provide a unique known key for each client station. This unique intermediate key is mixed with a 48-bit sequence number to create a per-packet key, which is then handed over to the WEP encryption engine as if it were composed of a 24-bit IV and 104-bit WEP key. The sequence number is required to increment on each packet, and any out-of-sequence packets are dropped, preventing IV-reuse attacks. The 48-bit sequence space means that a sequence number will not be reused on the order of 1000 years at today's data rates, so there are no repeated cipher streams to intercept. A sophisticated integrity checking mechanism is also included to guard against an attacker injecting slight variations of valid transmitted packets.

WPA addresses all the currently known attacks on WEP, though total security also depends on proper selection and implementation of algorithms within the 802.1x authentication process. It is certainly secure enough for home networks and for most enterprise/industrial implementations. At the time of this writing (mid-2004), the 802. 11i task group has approved an enhanced standard based on the Advanced Encryption Standard rather than the WEP RC4 algorithm, which will provide an adequate level of security for most uses.

Enterprises that have serious concerns about sensitive data can also implement end to end security through the use of VPNs and SSL web security. The advantage of this security approach is that protection against eavesdropping at any stage of the communications process, not just the wireless link, is obtained. However, VPNs are complex to set up and maintain and may not support roaming of the wireless device from one access point to another, particularly if the IP address of the mobile device undergoes a change due to the roaming process.

5.1.4 HiperLAN and HiperLAN 2

Some of the researchers who participated in defining the 802.11a standard were also active in similar European efforts: the resulting standards are known as HiperLAN. The HiperLAN 2 physical layer is very similar to that used in 802.11a: it is an OFDM system operating in the 5-GHz band, using an almost identical set of modulations and code rates to support nearly the same set of data rates. The physical layer is also compliant with ETSI requirements for dynamic frequency selection and power control.

However, the MAC layer is quite different. Instead of being based on the Ethernet standard, it is based on a totally different traffic approach, *asynchronous transfer mode* (ATM). ATM networking was developed by telephone service providers in the 1980s and 1990s in an attempt to provide a network that would support efficient transport of video, data, and voice, with control over the quality of service and traffic capacity assigned to various users. Native ATM is based on fixed 53-byte packets and virtual connections rather than the variable packets and global addressing used in IP networking over Ethernet. The HiperLAN MAC was constructed to provide a smooth interface to an ATM data network, by using fixed time slots assigned to stations by an access point acting as the central controller of the network.

ATM was originally envisioned as extending all the way from the WAN to the desktop but in practice has

seen very little commercial implementation beyond the data networks of large telecommunications service providers. Similarly, no commercial products based on HiperLAN have achieved significant distribution, and the author is not aware of any products in current distribution based on HiperLAN 2. It seems likely, particularly in view of recent FCC decisions bringing U.S. regulations into compliance with European regulations, that 802.11a products will achieve sufficient economies of scale to prevent the wide distribution of HiperLAN-based products in most applications. Although HiperLAN provides superior quality-of-service controls to 802.11, many major vendors for the key quality of service-sensitive service, video delivery, have transferred their attention to the ultrawideband (UWB) standardization activities in the 802.15.3a, discussed in the next section.

5.1.5 From LANs to PANs

WLANs are an attempt to provide a wireless extension of a computer network service. WPANs are intended for a somewhat different purpose and differ from their WLAN cousins as a consequence. The basic purpose of a WPAN is to replace a cable, not necessarily to integrate into a network attached to that cable. WPANs seek to replace serial and parallel printer cables, universal serial bus connections, and simple analog cables for speakers, microphones, and headphones with a single wireless digital data connection. Personal areas vary from person to person and culture to culture but are usually much smaller than a building, so PANs do not need to have the same range expected of WLANs. The cables being replaced are generally very inexpensive, so even more than WLANs, WPAN products must be inexpensive to build and use. Many WPAN products are likely to be included in portable battery-powered devices and should use power sparingly.

In this section we discuss three differing WPAN standards activities, all (at least today) taking place under the aegis of the 802.15 working group of IEEE. As is often the case, with the passing of time the definition of a category is stretched: what we've defined as a PAN for replacing low-rate cables is certainly evolving to replace higher rate cables and may be becoming a path to the universal home data/entertainment network that many companies have already unsuccessfully sought to popularize.

5.1.5.1 Bluetooth: Skip the danegeld, keep the dane

The first WPAN effort to achieve major visibility was the Bluetooth Special Interest Group, initiated in 1998 by Ericsson with support from Intel, IBM, Nokia, and Toshiba. The Bluetooth trademark is owned by Ericsson and licensed to users. Although the group clearly displayed more marketing savvy than all the IEEE working groups put together in its choice of a memorable moniker—the name refers to Harald "Bluetooth" Blätand, king of Denmark from about AD 940 to 985—the separation of this activity from the IEEE helped produce a standard that was not compatible in any way with 802.11 networks. Given the likelihood of collocated devices using the two standards, this is unfortunate. The Bluetooth standard is now also endorsed by the IEEE as 802.15.1.

To make life simple for users, the Bluetooth PAN was intended to operate with essentially no user configuration. Thus, Bluetooth networks are based on *ad hoc* discovery of neighboring devices and formation of networks. Simple modulation, low transmit power, low data rate, and modest sensitivity requirements all contribute to a standard that can be inexpensively implemented with minimal power requirements. Frequency hopping was used to allow multiple independent collocated networks. The network is based on synchronized time slots allocated by the master device, so that resources could be provided on a regular deterministic basis for voice and other time-sensitive traffic.

The architecture of Bluetooth networks is based on the *piconet* (Figure 5.1-22). In terminology also perhaps

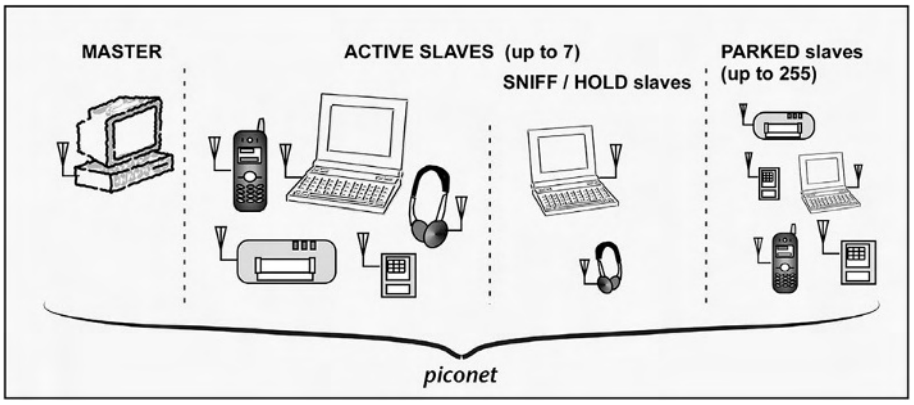


Figure 5.1-22 Bluetooth piconet.

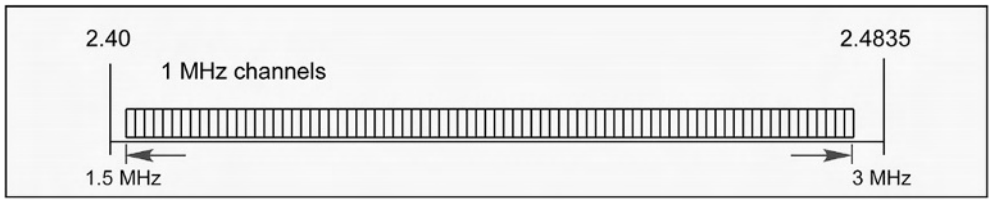


Figure 5.1-23 Bluetooth channelization.

harkening back to the time of King Blätand, a piconet consists of one device that acts as the *master* and a number of other devices that are *slaves*. Each piconet is identified by the FH pattern it uses. The hopping clock is set by the master device, and slaves must remain synchronized to the master to communicate. Provisions are made to permit devices both to discover nearby piconets and spontaneously form their own.

Only seven slave devices can be active at any given time. Some of the active devices can be placed in a power-saving *sniff* state, in which they only listen for their packets occasionally, or *hold* states, in which they do not actively participate in the piconet for some period of time before checking back in. Additional members of the piconet may be *parked:* in the parked state, slaves maintain synchronization with the master clock by listening to beacons periodically but do not otherwise participate until instructed to become active.

Piconets can physically overlap with one another with minimal interference, as they use distinct hopping schemes. A slave in one piconet could be a master in another. By having some devices participate in multiple piconets, they can be linked together to form *scatternets*. Devices within a piconet can exchange information on their capabilities, so that the user may use printers, cameras, and other peripherals without the need for manual configuration. Devices use *inquiry* and *paging* modes to discover their neighbors and form new piconets.

The Bluetooth/802.15.1 PHY layer operates in the 2.4-GHz ISM band, which is available at least partially throughout most of the world. The band in the United States is divided into 79 1-MHz channels (Figure 5.1-23), with modest guard bands at each end to minimize out-of-band emissions.

During normal operation, all members of the piconet hop from one frequency to another 1600 times per second. The dwell time of 625 μsec provides for 625 bit times between hops. When a device is in inquiry or paging mode, it hops twice as fast to reduce the time needed to find other devices.

The signal is modulated using *Gaussian minimum-shift keying*, GMSK (also known as Gaussian frequency-shift keying). GMSK can be regarded as either a frequency modulation or phase modulation technique. In each bit time, frequency is either kept constant or changed so as to shift the phase of the signal by π by the end of the period T (Figure 5.1-24). Because the amplitude remains constant during this process, the peak-to-average ratio of a GMSK signal is essentially 1, so such signals place modest requirements on transmitter and receiver linearity. GMSK is similar to BPSK, but BPSK systems have no such constraint on the trajectory taken between phase points and thus may have large amplitude variations during the transition from one symbol to another. Each GMSK symbol carries only a single bit, and thus 1 Msps = 1 Mbps of data transfer capacity.

Bluetooth radios come in three classes: class 1 can transmit up to 20 dBm (100 mW), class 2 is limited to 4 dBm (2.5 mW), and class 3 devices, the most common, operate at 1 mW (0dBm). Bluetooth receivers are required to achieve a bit error rate of 0.1% for a received signal at −70 dBm: that's a raw bit error rate of about 60% for 600-bit packets. These specifications are rather less demanding than the corresponding 802.11 requirements: recall that an 802.11 receiver is required to achieve a sensitivity of −80 dBm at 2 Mbps and typical transmit power is 30–100 mW. A link between two compliant class 3 devices has 30 dB less path loss available than a link between two 802.11 devices. The receiver noise and transmit power requirements are correspondingly undemanding: Bluetooth devices must be inexpensive to build.

An example of a link budget calculation is shown in Figure 5.1-25. We've assumed a transmit power of 1 dBm and modest antenna directivity (1 dB relative to an isotropic antenna [dBi]) and efficiency (80%) as the antennas

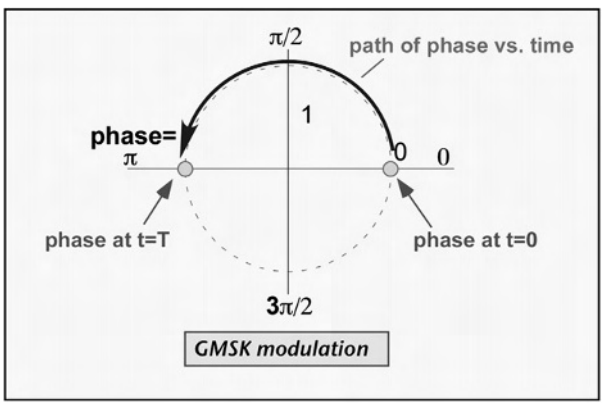


Figure 5.1-24 Gaussian minimum-shift keying.

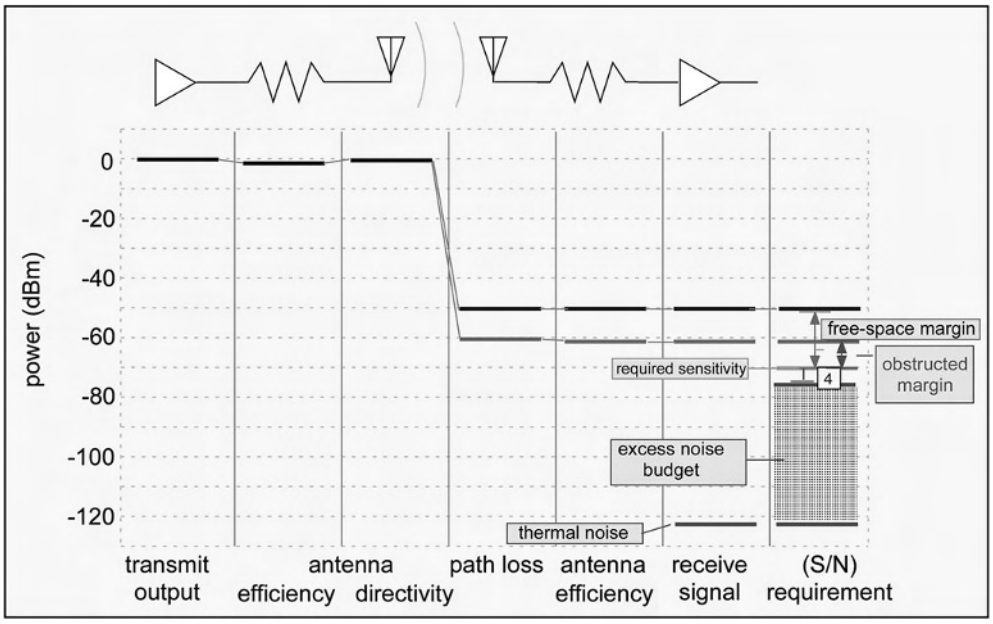


Figure 5.1-25 Bluetooth example link budget: 10 m, 0 dBm TX power.

are constrained to be quite small to fit into portable/handheld devices. At 3 m the free-space path loss is about 51 dB; we assumed that if a person is in the way of the direct path, an additional 10 dB of obstructed loss is encountered. If we assume a modest 4 dB (S/N) is required by the GMSK modulation, we find that we have room for 39 dB of excess noise from the receiver while still providing 10 dB of link margin in the presence of an obstruction. This is a huge noise figure; practical Bluetooth devices can do much better. The Bluetooth specifications are appropriate for inexpensive devices meant to communicate at moderate rates over ranges of a few meters.

A few words about the security of a Bluetooth link seem appropriate. Bluetooth links have two fundamental security advantages over 802.11 irrespective of any encryption technology. The first is a physical limit: a Bluetooth class 3 transmitter at 0 dBm is harder to intercept from afar than an 802.11 access point at 20-dBm transmit power. The second is related to the usage of the devices. Most 802.11 devices are connected to a local network of some kind, which provides many resources of interest to an attacker: servers containing possibly valuable data and a likely connection to the global Internet. Many Bluetooth devices may only provide access to a single data stream or a few neighbor devices. The attacker who intercepts data sent to a printer from a few meters away might just as well walk over to the printer and look at what's coming out. Finally, any given Bluetooth link is likely to be used less intensively and provides a lower peak data rate than an 802.11 link, so an attacker has a harder time gathering data to analyze.

That being said, Bluetooth provides security services that are in some ways superior to those in 802.11. Authentication and encryption are both provided and use distinct keys. The cipher stream generation uses the E0 sequence generator, which is at the present regarded as harder to attack than RC4. However, just like in 802.11, weaknesses are present due to the usage models. Keys for devices with a user interface are based on manually entered identifier (PIN) codes. Users won't remember long complex codes and don't do a very good job of choosing the codes they use at random, so links are likely to be vulnerable to brute force key-guessing attacks. Devices without a user interface have fixed PIN codes, obviously more vulnerable to attack. Devices that wish to communicate securely must exchange keys in a pairing operation; if an eavesdropper overhears this operation, they are in an improved position to attack the data.

Although there hasn't been as much publicity about cracking Bluetooth, as was the case with 802.11, it is prudent to use higher layer security (VPN or SSL, etc.) if large amounts of sensitive data are to be transferred over a Bluetooth link.

5.1.5.2 Enhanced PANs: 802.15.3

The 802.15.3 task group was created to enhance the capabilities of 802.15.1 (Bluetooth) while preserving the architecture, voice support, low cost of implementation, and range of the original. Data rates as high as 55 Mbps were targeted to allow applications such as streaming audio and video, printing high-resolution images, and transmitting presentations to a digital projector (Table 5.1-3). Because the work was now under the aegis of the IEEE, the group sought to create a PHY with some compatibility with existing 802.11 devices;

Table 5.1-3 802.15.3 Data Rates.

Modulation	Coding	Code rate	Data rate (Mbps)	Sensitivity (dBm)
QPSK	TCM	1/2	11	−82
DQPSK	None	–	22	−75
16QAM	TCM	3/4	33	−74
32QAM	TCM	4/5	44	−71
64QAM	TCM	2/3	55	−68

thus a symbol rate of 11 Msps was chosen, and the channels were to be compatible with the 5-MHz 802.11 channels. However, the group sought to use a slightly reduced bandwidth of 15 MHz, so that more nonoverlapping channels could be provided in the same area than 802.11 classic or b/g allow.

In July 2003, the 802.15.3 task group approved a new PHY layer standard that provides higher data rates while otherwise essentially using the Bluetooth MAC and protocol stack. The general approach is to use more sophisticated coding than was available a decade ago to allow more data to be transferred in less bandwidth. As with 802.11a and g, higher modulations are used to send more data per symbol; 32QAM uses the 16QAM constellation with the addition of partial rows and columns.

A novel coding approach, *trellis-coded modulation* (TCM), is used to achieve good performance at modest (S/N). In the conventional coding schemes we examined so far, each set of bits determines a modulation symbol (for example, a dibit determines a QPSK point) and the received voltage is first converted back to a digital value (an estimate of what the dibit was) and then used to figure out what the transmitted code word probably was. Coding is a separate and distinct operation from modulation. TCM mixes coding and modulation operations by dividing the constellation into subsets, within each of which the distance between signal points is large. The choice of subset is then made using a convolutional code on some of the input bits, whereas the choice of points within the subset is made using the remainder of the input bits uncoded. The subset points are widely separated and noise resistant, so uncoded selection works fine. The subsets are more difficult to distinguish from one another, but this choice is protected by the code, making it more robust to errors. The net result is typically around 3 dB of improved noise margin in the same bandwidth versus a conventional modulation.

The benefits of this approach are revealed in the spectral emission mask (Figure 5.1-26). The 802.15.3 signal is able to deliver the same data rate as 802.11g but fits completely within a 15-MHz window. The practical consequence is that one can fit five nonoverlapping channels into the U.S. ISM band, providing more flexibility in channel assignment and reducing interference. Some exposure to increased intersymbol interference and sensitivity to multipath may result from the narrower bandwidth, but because 802.15.3 is a PAN technology not intended to be used at long ranges, little performance limitation should result.

5.1.5.3 UWB PANs: A progress report

Although 55 Mbps may seem like a lot of bits, it is a marginal capacity for wireless transport of high-definition television signals (a technology that may finally reach mainstream deployment in this decade in conjunction with large-screen displays). Task group 802.15.3a was charged with achieving even higher data rates, at minimum 100 Mbps with a target of 400 Mbps, to support versatile short-range multimedia file transfer for homes and business applications.

To achieve such ambitious goals, the task group turned to a new approach, based again on recent FCC actions. In 2002 the FCC allowed a novel use of spectrum, *ultra-wideband transmission*. UWB radios are allowed to transmit right on top of spectrum licensed for other uses

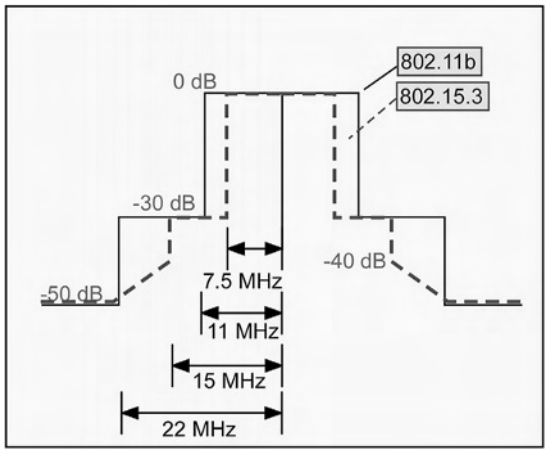


Figure 5.1-26 Comparison of 802.11b and 802.15.3 emission spectral mask.

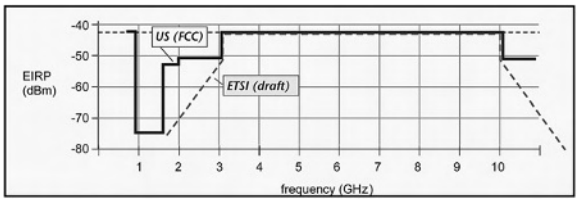


Figure 5.1-27 U.S. FCC and proposed ETSI limits for ultrawideband emitters (EIRP is equivalent isotropic radiated power).

(Figure 5.1-27). However, UWB radios operate under several restrictions designed to minimize interference with legacy users of the spectrum. First, the absolute power emitted at any band is restricted to less than −40 dBm, which is comparable with the emissions allowed by nonintentional emitters like computers or personal digital assistants. The restrictions are more stringent in the 1- to 3-GHz region to protect cellular telephony and global position satellite navigation. Finally, like the original requirements on the ISM band, UWB users are required to spread their signals over at least 500 MHz within the allowed band.

The wide bandwidth available for a UWB signal means that the peak data rate can be very high indeed. However, the limited total power implies that the range of transmission is not very large. As with any transmission method, UWB range can be extended at the cost of reduced data rate by using more complex coding methods to allow a tiny signal to be extracted from the noise.

The resulting trade-off between range and rate is shown in Figure 5.1-28. Here a full band transmission is one that exploits the whole 7-GHz bandwidth available under the FCC specs, whereas a subband transmission uses only the minimum 500-MHz bandwidth. It is clear that a subband approach can deliver 100 Mbps at a few meters, appropriate to a PAN application. As the demanded range increases, even the full UWB band cannot provide enough coding gain to substitute for transmit power, and the performance of the UWB transmitter falls below that of a conventional 802.11a system. UWB systems do not replace conventional WLAN systems but have ideal characteristics for WPAN applications.

At the time of this writing, two distinct proposals are under consideration by the 802.15.3a task group as

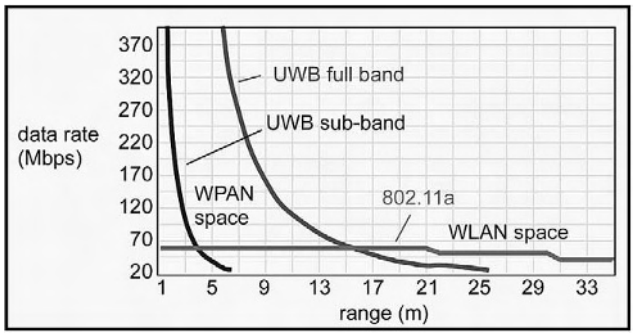


Figure 5.1-28 UWB range vs. data rate, shown for both full band (7 GHz) and subband (500 MHz) approaches (after Gandolfo, Wireless Systems Design Conference, 2003).

a WPAN standard. We first examine the direct-sequence CDMA proposal, based on 802.15−03 334/r3 due to Welborn et al., with some updates from /0137… r0. This proposal divides the available bandwidth into two chunks, a low band and a high band, avoiding the 5-GHz UNII region to minimize interference with existing and projected unlicensed products (Figure 5.1-29). The radio can operate using either the low band only, the high band only, or both bands, achieving data rates of 450, 900, or 1350 Mbps, respectively. Alternatively, the dual-band operating mode can be used to support *full-duplex* communications, in which the radio can transmit in one band and simultaneously receive in the other.

The transmissions use Hermitian pulses, the nomenclature referring to the pulse shape and consequent spectrum (Figure 5.1-30). The high band pulse is shown; a longer pulse is used for the low band.

The modulation techniques used are either BPSK or a variant of the biphase keying: *M-ary biorthogonal keying* (MBOK). In this scheme, the allowed symbols are specific sequences of positive and negative pulses; the sequences can be arranged in a hierarchy of mutually orthogonal symbols, in the sense that if they are multiplied together timewise and the sum taken (i.e., if they are correlated), the sum is 0 for distinct sequences. An example for $M = 4$ is shown in Figure 5.1-31. Each possible symbol is composed of four pulses, shown in cartoon form on the left. In the shorthand summary on the right, an inverted pulse is denoted "−" and a noninverted pulse

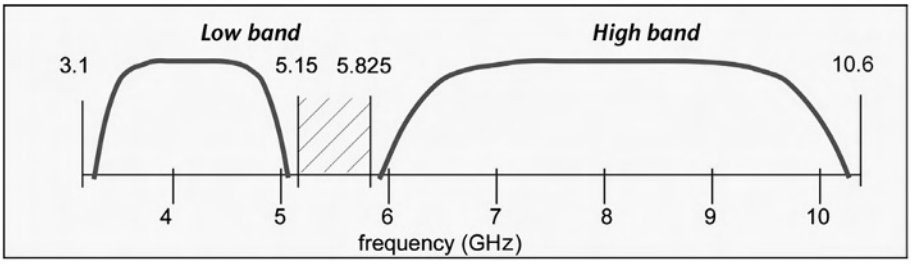


Figure 5.1-29 Pulse UWB band structure.

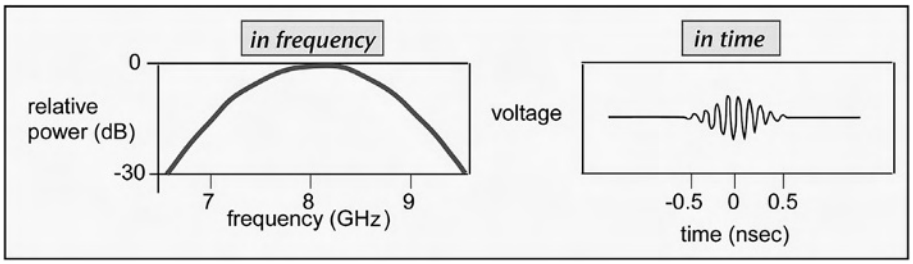


Figure 5.1-30 High-band modified Hermitian pulse.

is "+." Clearly, the first two symbols (00 and 01) are inversions of each other, as are the last two (10 and 11). Either symbol in the first group is orthogonal to either symbol in the second, for example, 00.10 = −1 +1 −1 +1=0, where we have assumed the product of a positive pulse and an inverted pulse to be −1.

The proposed radio architecture uses the MBOK symbols as chips in a CDMA system. Each data bit is multiplied by a 24- or 32-chip code. This scheme is similar to the Barker code used to multiply data bits in the 802.11 classic PHY, but unlike 802.11, in this case multiple codes are provided to enable simultaneous noninterfering use of the bands. Four independent 24-chip codes are provided, allowing four independent users of each band or eight total collocated piconets. A *rake receiver* can be used to separately add the contributions of sequences of chips arising from paths with different delays; a *decision feedback equalizer* is assigned to deal with interference in the data symbols resulting from multipath. Both convolution and block code options are provided; the codes may be *concatenated* for improved performance at the cost of computational complexity.

Link performance at a transmit-receive distance of 4 m and data rate of 200 Mbps is summarized in Table 5.1-4. Note that the transmit power, −10 dBm, is 10 dB less than the lowest Bluetooth output and about 25 dB less than a typical 802.11 transmitter. The required bit energy over noise, E_b/N_o, is comparable with what we have encountered in conventional BPSK or QPSK modulation. Allowing for a couple of decibels of "implementation loss" (antenna limitations, cable losses, etc.) and a noise figure of 6.6 dB, the link should still achieve a respectable −75 dBm sensitivity, allowing some margin at a few meters range.

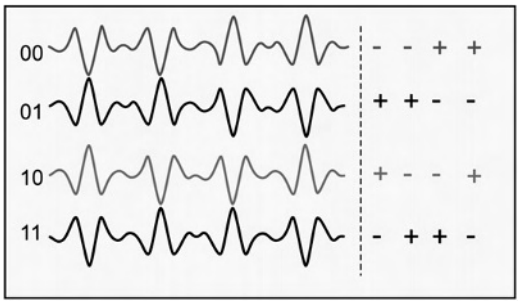


Figure 5.1-31 4-Ary biorthogonal symbols.

The second current proposal is based on multiband OFDM, as described in IEEE 802.15–03/267r2, r6 Batra et al., with some updates from 802.15–04/0122r4 of March 2004. In this case, the available UWB bandwidth is partitioned into 528-MHz bands (recall that 500 MHz is the smallest bandwidth allowed by the FCC for an UWB application), as shown in Figure 5.1-32. The bands are assembled into five groups. Group 1 devices are expected to be introduced first, with more advanced group 2–5 devices providing improved performance. All devices are required to support group 1 bands; the higher bands are optional.

The implementation requires a 128-point inverse FFT to create subcarriers spanning slightly more than 500MHz of bandwidth. To reduce the computational demands, QPSK modulation on each subcarrier is used, instead of the higher QAM modulations used in 802.11 a/g. Furthermore, for data rates less than 80Mbps, the subcarrier amplitudes are chosen to be conjugate symmetric around the carrier. That is, the *n*th subcarrier at

Table 5.1-4 Link performance, DS-CDMA UWB at 200 mbps.

Parameter	Value
Data rate	200 Mbps
TX power	–10 dBm
Path loss	56 dB (4m)
RX power	–66 dBm
Noise/bit	–91 dBm
RX noise figure	6.6 dB
Total noise	–84.4 dBm
Required E_b/N_0	6.8 dB
Implementation loss	2.5 dB
Link margin	8.7 dB
RX sensitivity	–75 dBm

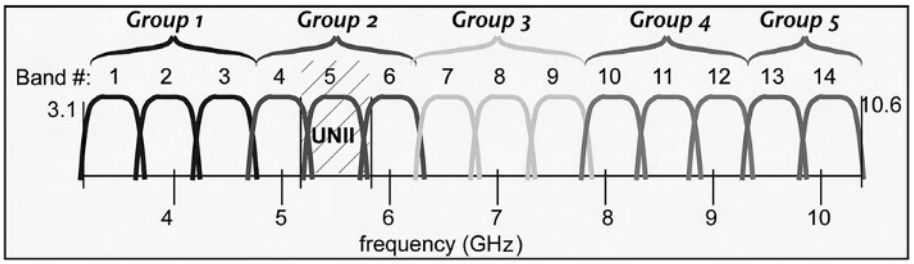


Figure 5.1-32 Band definitions for OFDM UWB proposal.

positive frequency (relative to the carrier) is the complex conjugate of the nth subcarrier at negative frequency. A signal composed of complex conjugate frequencies produces a pure real voltage (this is just a fancy way of reminding us that $e^{i\omega t} + e^{-i\omega t} = 2\cos(\omega t)$). This means that the transmitter and receiver don't need separate branches to keep track of the in-phase and quadrature (I and Q) components of the signal and can be simplified.

Successive OFDM symbols are sent on different bands within a band group. Within each band group 1–4 there are four hopping patterns, each six symbols long, allowing for four collocated piconets. (Group 5 supports only two hopping patterns.) Thus, a total of 18 collocated piconets is possible.

The implementation parameters for the 200 Mbps are summarized in Table 5.1-5. A convolutional code with rate 5/8 (8 output bits for 5 input bits) is used. Of the 128 subcarriers, 22 are used as pilot tones and guard tones and 6 are not used at all. The resulting symbol is about 1/4 of a microsecond long. The cyclic prefix is no longer included, because this prefix adds a periodicity to the signal, giving rise to spectral lines which reduce the amount of power that can be transmitted. Instead, an equivalent amount of zero padding is added in frequency space (*zero-padded OFDM*). Each data symbol is spread by being sent twice, in separate frequency subbands. A brief guard interval is provided between symbols to allow the radio to change subbands. The use of separate subbands allows multiple collocated piconets.

In Table 5.1-6 we compare link performance for the two proposals for a data rate of 200 Mbps at 4 m. OFDM has a slight advantage in noise tolerance, giving it modestly improved sensitivity. However, as we noted in connection with the 802.11b PHY, small distinctions in radio performance are likely to be less important than cost and ease of use in determining commercial viability.

Recent modifications (March 2004) to the direct-sequence proposal have changed the MBOK signaling and proposed a common signaling mechanism between the multiband OFDM and direct sequence approaches. The

Table 5.1-5 Wideband OFDM parameters, 200 Mbps rate.

Parameter	Value
Data rate	200 Mbps
Modulation constellation	OFDM/QPSK
FFT size	128 tones
Coding rate	R = 5/8
Spreading rate	2
Pilot/guard tones	22
Data tones	100
Information length	242.4 nsec
Padded prefix	60.6 nsec
Guard interval	9.5 nsec
Symbol length	312.5 ns
Channel bit rate	640 Mbps

Table 5.1-6 Link performance, 200 Mbps.

Parameter	Wideband OFDM	DS-CDMA
Information data rate	200 Mbps	200 Mbps
Average TX power	–10 dBm	–10 dBm
Total path loss	56 dB (4m)	56 dB(4m)
Average RX power	–66 dBm	–66 dBm
Noise power per bit	–91.0 dBm	–91 dBm
Complementary metal-oxide semiconductor RX noise figure	6.6 dB	6.6 dB
Total noise power	–84.4 dBm	–84.4 dBm
Required E_b/N_o	4.7 dB	6.8 dB
Implementation loss	2.5 dB	2.5 dB
Link margin	10.7 dB	8.7 dB
RX sensitivity level	–77.2 dBm	–75 dBm

selection process appears to be favoring the multiband OFDM proposal, but it has not yet been approved. At the time of this writing, it is not known whether either of these proposals will be approved by the IEEE as standards or whether the resulting standard will enjoy any success in the marketplace. However, even if never deployed, these proposals provide an indication of the likely approaches for future high-rate short-range radio technologies.

5.1.6 Capsule summary: Chapter 5.1

WLANs are constructed to extend LANs and share some of the properties of their wired forebears. However, the wireless transition places requirements on any WLAN protocol for connection management, medium allocation, error detection and correction, and additional link security. IEEE 802.11b WLANs provided a good combination of range and data rate at low cost and easy integration with popular Ethernet-based wired LANs; the successful WECA consortium also provided assurance of interoperability between vendors. These factors made 802.11b Wi-Fi networks popular despite limited support for time-sensitive traffic and porous link security. Enhancements of the 802.11b standard to higher data rates (802.11g) and different bands (802.11a) provide an evolution path to support more users and applications; WPA and eventually 802.11i enable a more robust encryption and authentication environment for protecting sensitive information.

Contemporaneous with the evolution of WLANs, WPANs have been developed to replace local cables. For historical and technical reasons, a distinct PAN architecture has evolved, using master/slave self-organizing piconets. WPAN radios are targeted for short-range communications and require modest transmit power and limited receiver sensitivity. Support for time-sensitive traffic is provided in the design. WPANs are evolving toward higher data rates and in a possible UWB implementation may provide the basis for the long-envisioned convergence of the various home entertainment/ information media into a single short-range wireless network.

Further reading

802.11 Networks

802.11 Networks: The Definitive Guide, Matthew Gast (cited in chapter 1): *The best overall introduction. Goes far beyond what we have provided here in examining framing and network management.*

Wireless Local Area Networks, Benny Bing (ed.), Wiley-Interscience, 2002: *Because this is a collection of chapters by different authors from different organizations, the quality of the chapters varies considerably, but the book covers a lot of ground. Chapter 2 (Heegard et al.) provides a very nice discussion of the performance of 802.11b codes.*

How Secure is Your Wireless Network?, Lee Barken, Prentice-Hall, 2004: *Provides a somewhat more detailed discussion of the flaws in WEP and a practical introduction to the use of WPA, EAP, and VPN setup.*

WEP Attacks

"Weaknesses in the Key Scheduling Algorithm of RC4," S. Fluhrer, I. Mantin, and A. Shamir, Selected Areas in Cryptography conference, Toronto, Canada, 16–17 August 2001

"Using the Fluhrer, Mantin and Shamir Attack to Break a WEP," A. Stubblefield, J. Ioannidis, and A. Rubin, Network and Distributed Security Symposium Conference, 2002

Bluetooth

Bluetooth Revealed, Brent Miller and Chatschik Bisdikian, Prentice-Hall, 2001: *Nice description of how the Bluetooth protocols work to provide discovery, association, and network traffic management.*

Trellis-Coded Modulations

Digital Modulations and Coding, Stephen Wilson (cited in chapter 2): *See section 6.6 for a more detailed examination of how TCM works.*

Standards

http://www.ieee802.org/ is the home page for all the 802 standards activities. At the time of this writing, 802 standards are available for free download as Adobe Acrobat PDF files 6 months after they are approved. Many working documents, such as those cited in the UWB description in section 5.3, are also available on the respective task group web pages. Draft standards that are not yet available for free download can be purchased from the IEEE store (but beware of awkward digital-rights-management limitations on these files, at least as of 4Q 2003).

http://www.etsi.org/SERVICES_PRODUCTS/FREESTANDARD/HOME.HTM provides access to ETSI standards as PDF downloads. The HiperLAN standards are described in a number of ETSI documents; three useful ones to start with are as follows:

ETSI TS 101 475 v 1.3.1 (2001): "Broadband Radio Access Networks… HiperLAN 2 PHYSICAL LAYER"

TR 101 031 v 1.1.1 (1997) "Radio Equipment and Systems… (HiperLAN)…" and

ETSI EN 301 893 v 1.2.3 (2003) "Broadband Radio Access Networks… 5 GHz High Performance RLAN"

http://www.bluetooth.com/ is the home page for the Bluetooth Special Interest Group (SIG), but at the time of this writing it appears that standardization activity is not included at this web site.

This page intentionally left blank

Chapter 5.2

5.2

Short-range wireless applications and technologies

Alan Bensky

An important factor in the widespread penetration of short-range devices into the office and the home is the basing of the most popular applications on industry standards. In this chapter, we take a look at some of these standards and the applications that have emerged from them. Those covered pertain to HomeRF, Wi-Fi, HIPERLAN/2, Bluetooth, and Zigbee. In order to be successful, a standard has to be built so that it can keep abreast of rapid technological advancements by accommodating modifications that don't obsolete earlier devices that were developed to the original version. A case in point is the competition between the WLAN (wireless local area network) standard that was developed by the HomeRF Working Group based on the SWAP (shared wireless access protocol) specification, and IEEE specification 802.11, commonly known as Wi-Fi. The former used frequency-hopping spread-spectrum exclusively, and although some increase of data rate was provided for beyond the original 1 and 2 Mbps, it couldn't keep up with Wi-Fi, which incorporated new bandwidth efficient modulation methods to increase data rates 50-fold while maintaining compatibility with first generation DSSS terminals. Other reasons why HomeRF lost out to Wi-Fi are given below.

Many of the new wireless short-range systems are designed for operation on the 2.4 GHz ISM band, available for license-free operation in North America and Europe, as well as virtually all other regions in the world. Most systems have provisions for handling errors due to interference, but when the density of deployment of one or more systems is high, throughput, voice intelligibility, or quality of service in general is bound to suffer. We will look at some aspects of this problem and methods for solving it in relation to Bluetooth and Wi-Fi.

A relatively new approach to short-range communications with unique technological characteristics is ultra-wideband (UWB) signal generation and detection. UWB promises to add applications and users to short-range communication without impinging on present spectrum use. Additionally, it has other attributes including range finding and high power efficiency that are derived from its basic principles of operation. We present the main features of UWB communication and an introduction to how it works.

5.2.1 Wireless local area networks (WLAN)

One of the hottest applications of short-range radio communication is wireless local area networks. While the advantage of a wireless versus wired LAN is obvious, the early versions of WLAN had considerably inferior data rates so conversion to wireless was often not worthwhile, particularly when portability is not an issue. However, advanced modulation techniques have allowed wireless throughputs to approach and even exceed those of wired networks, and the popularity of highly portable laptop and handheld computers, along with the decrease in device prices, have made computer networking a common occurrence in multi-computer offices and homes.

There are still three prime disadvantages to wireless networks as compared to wired: range limitation, susceptibility to electromagnetic interference, and security. Direct links may be expected to perform at a top range of 50 to 100 meters depending on frequency band and surroundings. Longer distances and obstacles will reduce data throughput. Greater distances between network participants are achieved by installing additional access points to bridge remote network nodes. Reception of radio signals may be interfered with by other services operating on the same frequency band and in the same vicinity. Wireless transmissions are subject to

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2004 Elsevier Inc. All rights reserved

eavesdropping, and a standardized security implementation in Wi-Fi called WEP (wired equivalent privacy), has been found to be breachable with relative ease by persistent and knowledgeable hackers. More sophisticated encryption techniques can be incorporated, although they may be accompanied by reduction of convenience in setting up connections and possibly in performance.

Various systems of implementation are used in wireless networks. They may be based on an industrial standard, which allows compatibility between devices by different manufacturers, or a proprietary design. The latter would primarily be used in a special purpose network, such as in an industrial application where all devices are made by the same manufacturer and where performance may be improved without the limitations and compromises inherent in a widespread standard.

The HomeRF Working Group

The HomeRF Working Group was established by prominent computer and wireless companies that joined together to establish an open industry specification for wireless digital communication between personal computers and consumer electronic devices anywhere in and around the home. It developed the SWAP specification—Shared Wireless Access Protocol, whose major application was setting up a wireless home network that connects one or more computers with peripherals for the purposes of sharing files, modems, printers, and other electronic devices, including telephones. In addition to acting as a transparent wire replacement medium, it also permitted integration of portable peripherals into a computer network. The originators expected their system to be accepted in the growing number of homes that have two or more personal computers.

Following are the main system technical parameters:

• Frequency-hopping network	50 hops per second
• Frequency range	2.4 GHz ISM band
• Transmitter power	100 milliwatt
• Data rate	1 Mbps using 2FSK modulation 2 Mbps using 4FSK modulation
• Range	Covers typical home and yard
• Supported stations	Up to 127 devices per network
• Voice connections	Up to 6 full-duplex conversations
• Data security	Blowfish encryption algorithm (over 1 trillion codes)
• Data compression	LZRW3-A (Lempel-Ziv) algorithm
• 48-bit network ID	Enables concurrent operation of multiple co-located networks

The HomeRF Working Group ceased activity early in 2003. Several reasons may be cited for its demise. Reduction in prices of its biggest competitor, Wi-Fi, all but eliminated the advantage HomeRF had for home networks—low cost. Incompatibility with Wi-Fi was a liability, since people who used their Wi-Fi equipped laptop computer in the office also needed to use it at home, and a changeover to another terminal accessory after work hours was not an option. If there were some technical advantages to HomeRF, support of voice and connections between peripherals for example, they are becoming insignificant with the development of voice interfaces for Wi-Fi and the introduction of Bluetooth.

Wi-Fi

Wi-Fi is the generic name for all devices based on the IEEE specification 802.11 and its derivatives. It is promoted by the Wi-Fi Alliance that also certifies devices to ensure their interoperability. The original specification is being continually updated by IEEE working groups to incorporate technical improvements and feature enhancements that are agreed upon by a wide representation of potential users and industry representatives. 802.11 is the predominant industrial standard for WLAN and products adhering to it are acceptable for marketing all over the world.

802.11 covers the data link layer of lower-level software, the physical layer hardware definitions, and the interfaces between them. The connection between application software and the wireless hardware is the MAC (medium access control). The basic specification defines three types of wireless communication techniques: DSSS (direct sequence spread spectrum), FSSS (frequency-hopping spread spectrum) and IR (infra-red). The specification is built so that the upper application software doesn't have to know what wireless technique is being used—the MAC interface firmware takes care of that. In fact, application software doesn't have to know that a wireless connection is being used at all and mixed wired and wireless links can coexist in the same network.

Wireless communication according to 802.11 is conducted on the 2.400 to 2.4835 GHz frequency band that is authorized for unlicensed equipment operation in the United States and Canada and most European and other countries. A few countries allow unlicensed use in only a portion of this band. A supplement to the original document, 802.11b, adds increased data rates and other features while retaining compatibility with equipment using the DSSS physical layer of the basic specification. Supplement 802.11a specifies considerably higher rate operation in bands of frequencies between 5.2 and 5.8 GHz. These data rates were made available on the 2.4 GHz band by 802. 11g that has downward compatibility with 802.11b.

Network architecture

Wi-Fi architecture is very flexible, allowing considerable mobility of stations and transparent integration with wired IEEE networks. The transparency comes about because upper application software layers (see below) are not dependent on the actual physical nature of the communication links between stations. Also, all IEEE LAN stations, wired or wireless, use the same 48-bit addressing scheme so an application only has to reference source and destination addresses and the underlying lower-level protocols will do the rest.

Three Wi-Fi network configurations are shown in Figures 5.2-1 through 5.2-3. Figure 5.2-1 shows two unattached basic service sets (BSS), each with two stations (STA). The BSS is the basic building block of an 802.11 WLAN. A station can make *ad hoc* connections with other stations within its wireless communication range but not with those in another BSS that is outside of this range. In order to interconnect terminals that are not in direct range one with the other, the distributed system shown in Figure 5.2-2 is needed. Here, terminals that are in range of a station designated as an access point (AP) can communicate with other terminals not in direct range but who are associated with the same or another AP. Two or more such access points communicate between themselves either by a wireless or wired medium, and therefore data exchange between all terminals in the network is supported. The important thing here is that the media connecting the STAs with the APs, and connecting the APs among themselves are totally independent.

A network of arbitrary size and complexity can be maintained through the architecture of the extended service set (ESS), shown in Figure 5.2-3. Here, STAs have full mobility and may move from one BSS to another while remaining in the network. Figure 5.2-3 shows another element type—a portal. The portal is a gateway between the WLAN and a wired LAN. It connects the medium over which the APs communicate to the medium of the wired LAN—coaxial cable or twisted pair lines, for example.

In addition to the functions Wi-Fi provides for distributing data throughout the network, two other important services, although optionally used, are provided.

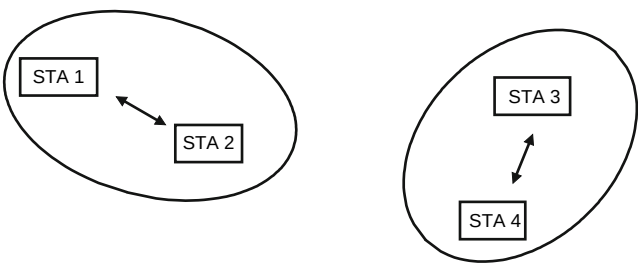


Figure 5.2-1 Basic service set.

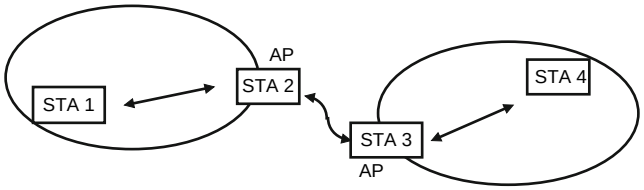


Figure 5.2-2 Distribution system and access points.

They are authentication and encryption. Authentication is the procedure used to establish the identity of a station as a member of the set of stations authorized to associate with another station. Encryption applies coding to data to prevent an eavesdropper from intercepting it. 802.11 details the implementation of these services in the MAC. Further protection of confidentiality may be provided by higher software layers in the network that are not part of 802.11.

The operational specifics of WLAN are described in IEEE 802.11 in terms of defined protocols between lower-level software layers. In general, networks may be described by the communication of data and control between adjacent layers of the Open System Interconnection Reference Model (OSI/RM), shown in Figure 5.2-4, or the peer-to-peer communication between like layers of two or more terminals in the network. The bottom layer, physical, represents the hardware connection with the transmission medium that connects the terminals of the network—cable modem, radio transceiver and antenna, infrared transceiver, or power line transceiver, for example. The software of the upper layers is wholly independent of the transmission medium and in principle may be used unchanged no matter what the nature of the medium and the physical connection to it. IEEE 802.11 is concerned only with the two lowest layers, physical and data link.

IEEE 802.11 prescribes the protocols between the MAC sublayer of the data layer and the physical layer, as well as the electrical specifications of the physical layer. Figure 5.2-5 illustrates the relationship between the physical and MAC layers of several types of networks with upper-layer application software interfaced through a commonly defined logical link control (LLC) layer. The LLC is common to all IEEE local area networks and is independent of the transmission medium or medium access method. Thus, its protocol is the same for wired local area networks and the various types of wireless networks. It is described in specification ANSI/IEEE standard 802.2.

The Medium Access Control function is the brain of the WLAN. Its implementation may be as high-level digital logic circuits or a combination of logic and a microcontroller or a digital signal processor. IEEE 802.11 and its supplements, (which may be generally designated 802.1 1x), prescribe various data rates, media (radio

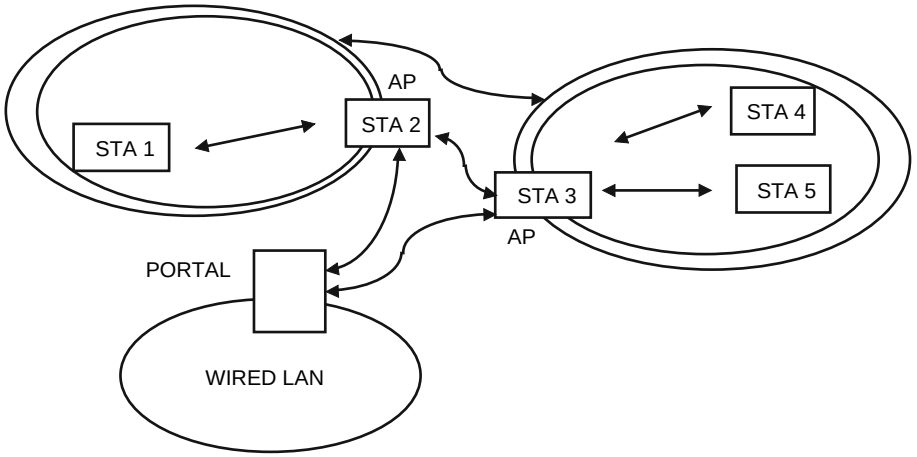


Figure 5.2-3 Extended service set.

waves or infrared), and modulation techniques (FHSS, DSSS, CCK, ODFM). These are the principle functions of the MAC:

- Frame delimiting and recognition
- Addressing of destination stations
- Transparent transfer of data, including fragmentation and defragmentation of packets originating in upper layers
- Protection against transmission error
- Control of access to the physical medium
- Security services—authentication and encryption.

An important attribute of any communications network is the method of access to the medium. Specification 802.11 prescribes two possibilities: DCF (distributed coordination function) and PCF (point coordination function).

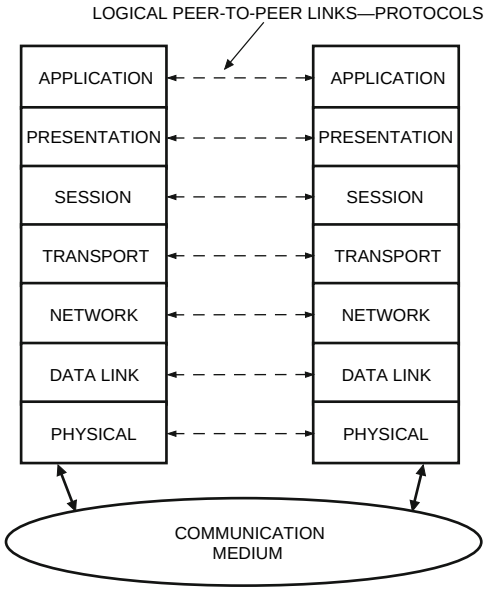


Figure 5.2-4 Open system interconnection reference model.

The fundamental access method in IEEE 802.11 is the DCF, more widely known as CSMA/CA (carrier sense multiple access with collision avoidance). It is based on a procedure during which a station wanting to transmit may do so only after listening to the channel and determining that it is not busy. If the channel is busy, the station must wait until the channel is idle. In order to minimize the possibility of collisions when more than one station wants to transmit at the same time, each station waits a random time-period, called a back off interval, before transmitting, after the channel goes idle. Figure 5.2-6 shows how this method works.

The figure shows activity on a channel as it appears to a station that is attempting to transmit. The station may start to transmit if the channel is idle for a period of at least a duration of DIFS (distributed coordination function interframe space) since the end of any other transmission (section 1 of the figure). However, if the channel is busy, as shown in section 2 of the figure, it must defer access and enter a back off procedure. The station waits until the channel is idle, and then waits an additional period of DIFS. Now it computes a time-period called a back off window that equals a pseudo-random number multiplied by constant called the "slot time." As long as the channel is idle, as it is in section 3 of the figure, the station may transmit its frame at the end of the back off window, section 4. During every slot time of the back off window the station senses the channel, and if it is busy, the counter that holds the remaining time of the back off window is frozen until the channel becomes idle and the back off counter resumes counting down.

Actually, the back off procedure is not used for every access of the channel. For example, acknowledgement transmissions and RTS and CTS transmissions, (see below), do not use it. Instead, they access the channel after an interval called SIFS (short interframe space) following the transmission to which they are responding. SIFS is shorter than DIFS, so other stations waiting to

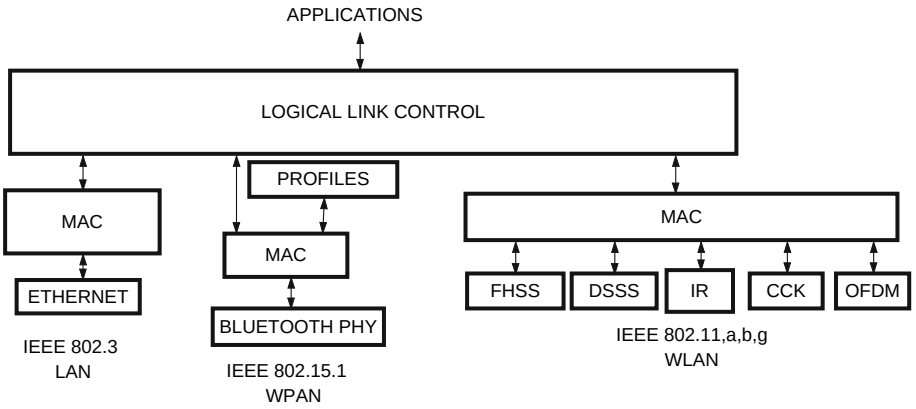


Figure 5.2-5 Data link and physical layers (PHY).

transmit cannot interfere since they have to wait a longer time, after the previous transmission, and by then the channel is already occupied.

In waiting for a channel to become idle, a transmission contender doesn't have to listen continuously. When one hears another station access the channel, it can interpret the frame length field that is transmitted on every frame. After taking into account the time of the acknowledgement transmission that replies to a data transmission, the time that the channel will become idle is known even without physically sensing it. This is called a virtual carrier sense mechanism.

The procedure shown in Figure 5.2-6 may not work well under some circumstances. For example, if several stations are trying to transmit to a single access point, two or more of them may be positioned such that they all are in range of the access point but not of each other. In this case, a station sensing the activity of the channel may not hear another station that is transmitting on the same network. A refinement of the described CSMA/SA procedure is for a station thinking the channel is clear to send a short RTS (request to send) control frame to the AP. It will then wait to receive a CTS (clear to send) reply from the AP, which is in range of all contenders for transmission, before sending its data transmission. If the originating station doesn't hear the CTS it assumes the channel was busy and so it must try to access the channel again. This RTS/CTS procedure is also effective when not all stations on the network have compatible modulation facilities for high rate communication and one station may not be able to detect the transmission length field of another. RTS and CTS transmissions are always sent at a basic rate that is common to all participants in the network.

The PCS is an optional access method that uses a master-slave procedure for polling network members. An AP station assumes the role of master and distributes timing and priority information through beacon management transmissions, thus creating a contention free access method. One use of the PCS is for voice communications, which must use regular time slots and will not work in a random access environment.

Physical layer

The discussion so far on the services and the organization of the WLAN did not depend on the actual type of wireless connection between the members of the network. Specification 802.11 and its additions specify various bit rates, modulation methods, and operating frequency channels, on two frequency bands, which we discuss in this section.

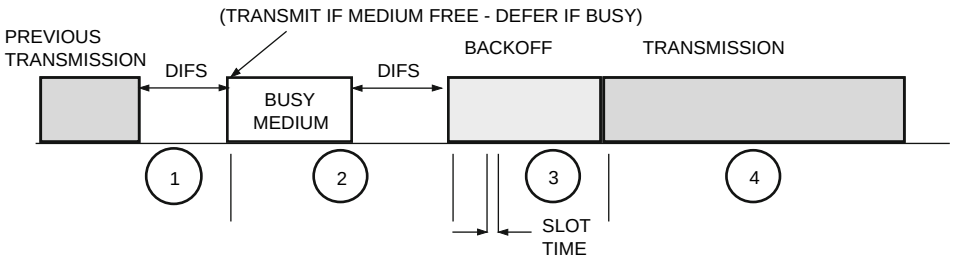


Figure 5.2-6 CSMA/CA access method.

IEEE 802.11 basic

The original version of the 802.11 specification prescribes three different air interfaces, each having two data rates. One is infrared and the others are based on frequency-hopping spread spectrum (FHSS) and direct-sequence spread-spectrum, each supporting raw data rates of 1 and 2 Mbps. Below is a short description of the IR and FHSS links, and a more detailed review of DSSS.

Infrared PHY

Infrared communication links have some advantages over radio wave transmissions. They are completely confined within walled enclosures and therefore eavesdropping concerns are greatly relieved, as are problems from external interference. Also, they are not subject to intentional radiation regulations. The IEEE 802.11 IR physical layer is based on diffused infrared links, and the receiving sensor detects radiation reflected off ceilings and walls, making the system independent of line-of-site. The range limit is on the order of 10 meters. Baseband pulse position modulation is used, with a nominal pulse width of 250 nsec. The IR wavelength is between 850 and 950 nM. The 1 Mbps bit rate is achieved by sending symbols representing 4 bits, each consisting of a pulse in one of 16 consecutive 250 nsec slots. This modulation method is called 16-PPM. Optional 4-PPM modulation, with four slots per two-bit symbol, gives a bit rate of 2 Mbps.

Although part of the original IEEE 802.11 specification and having what seems to be useful characteristics for some applications, products based on the infrared physical layer for WLAN have generally not been commercially available. However, point-to-point, very short-range infrared links using the IrDA (Infrared Data Association) standard are very widespread (reputed to be in more than 300 million devices). These links work reliably line-of-site at one meter and are found, for example, in desktop and notebook computers, handheld PC's, printers, cameras and toys. Data rates range from 2400 Bps to 16 Mbps. Bluetooth devices will take over some of the applications but for many cases IrDA imbedding will still have an advantage because of its much higher data rate capability.

FHSS PHY

While overshadowed by the DSSS PHY, acquaintance with the FHSS option in 802.11 is still useful since products based on it may be available. In FHSS WLAN, transmissions occur on carrier frequencies that hop periodically in pseudo-random order over almost the complete span of the 2.4 GHz ISM band. This span in North America and most European countries is 2.400 to 2.4835 GHz, and in these regions there are 79 hopping carrier frequencies from 2.402 to 2.480 GHz. The dwell on each frequency is a system-determined parameter, but the recommended dwell time is 20 msec, giving a hop rate of 50 hops per second. In order for FHSS network stations to be synchronized, they must all use the same pseudo-random sequence of frequencies, and their synthesizers must be in step, that is, they must all be tuned to the same frequency channel at the same time. Synchronization is achieved in 802.11 by sending the essential parameters—dwell time, frequency sequence number, and present channel number—in a frequency parameter set field that is part of a beacon transmission (and other management frames) sent periodically on the channel. A station wishing to join the network can listen to the beacon and synchronize its hop pattern as part of the network association procedure.

The FHSS physical layer uses GFSK (Gaussian frequency shift keying) modulation, and must restrict transmitted bandwidth to 1 MHz at 20 dB down (from peak carrier). This bandwidth holds for both 1 Mbps and 2 Mbps data rates. For 1 Mbps data rate, nominal frequency deviation is ± 160 kHz. The data entering the modulator is filtered by a Gaussian (constant phase delay) filter with 3 dB bandwidth of 500 kHz. Receiver sensitivity must be better than -80 dBm for a 3% frame error rate.

In order to keep the same transmitted bandwidth with a data rate of 2 Mbps, four-level frequency shift-keying is employed. Data bits are grouped into symbols of two bits, so each symbol can have one of four levels. Nominal deviations of the four levels are ± 72 kHz and ± 216 kHz. A 500 kHz Gaussian filter smoothes the four-level 1 Megasymbols per second at the input to the FSK modulator. Minimum required receiver sensitivity is -75 dBm.

Although development of Wi-Fi for significantly increased data rates has been along the lines of DSSS, FHSS does have some advantageous features. Many more independent networks can be collocated with virtually no mutual interference using FHSS than with DSSS. As we will see later, only three independent DSSS networks can be collocated. However, 26 different hopping sequences (North America and Europe) in any of three defined sets can be used in the same area with low probability of collision. Also, the degree of throughput reduction by other 2.4 GHz band users, as well as interference caused to the other users is lower with FHSS. FHSS implementation may at one time also have been less expensive. However, the updated versions of 802.11—specifically 802.11a, 802.11b, and 802.11g—have all based their methods of increasing data rates on the broadband channel characteristics of DSSS in 802.11, while being downward compatible with the

1 and 2 Mbps DSSS modes (except for 802.11a which operates on a different frequency band).

DSSS PHY

The channel characteristics of the direct sequence spread spectrum physical layer in 802.11 are retained in the high data rate updates of the specification. This is natural, since systems based on the newer versions of the specification must retain compatibility with the basic 1 and 2 Mbps physical layer. The channel spectral mask is shown in Figure 5.2-7, superimposed on the simulated spectrum of a filtered 1 Mbps transmission. It is 22 MHz wide at the −30 dB points. Fourteen channels are allocated in the 2.4 GHz ISM band, whose center frequencies are 5 MHz apart, from 2.412 GHz to 2.484 GHz. The highest channel, number fourteen, is designated for Japan where the allowed band edges are 2.471 GHz and 2.497 GHz. In the US and Canada, the first eleven channels are used. Figure 5.2-8 shows how channels one, six and eleven may be used by three adjacent independent networks without co-interference. When there are no more than two networks in the same area, they may choose their operating channels to avoid a narrow-band transmission or other interference on the band.

In 802.11 DSSS, a pseudo-random bit sequence phase modulates the carrier frequency. In this spreading sequence, bits are called chips. The chip rate is 11 megachips per second (Mcps). Data is applied by phase modulating the spread carrier. There are eleven chips per data symbol. The chosen pseudo-random sequence is a Barker sequence, represented as 1,–1,1,1,–1,1,1,1,–1, –1,–1. Its redeeming property is that it is optimally

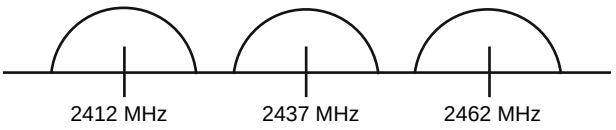


Figure 5.2-8 DSSS non-interfering channels.

detected in a receiver by a matched filter or correlation detector. Figure 5.2-9 is one possible implementation of the modulator. The DSSS PHY specifies two possible data rates—1 and 2 Mbps. The differential encoder takes the data stream and produces two output streams at 1 Mbps that represent changes in data polarity from one symbol to the next. For a data rate of 1 Mbps, differential binary phase shift keying is used. The input data rate of 1 Mbps results in two identical output data streams that represent the changes between consecutive input bits. Differential quadrature phase shift keying handles 2 Mbps of data. Each sequence of two input bits creates four permutations on two outputs. The differential encoder outputs the differences from symbol to symbol on the lines that go to the inputs of the exclusive OR gates shown in Figure 5.2-9. The outputs on the *I* and Q lines are the Barker sequence of 11 Mcps inverted or sent straight through, at a rate of 1 Msps, according to the differentially encoded data at the exclusive OR gate inputs. These outputs are spectrum shifted to the RF carrier frequency (or an intermediate frequency for subsequent up-conversion) in the quadrature modulator.

Reception of DSSS signals is represented in Figure 5.2-10. The downconverted *I* and Q signals are applied to matched filters or correlation detectors. These

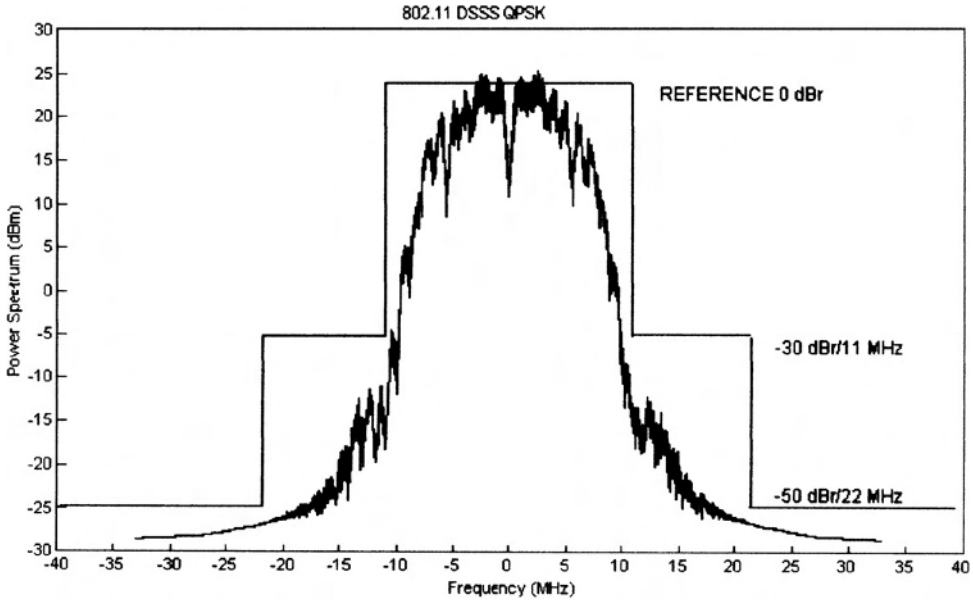


Figure 5.2-7 802.11 DSSS spectral mask.

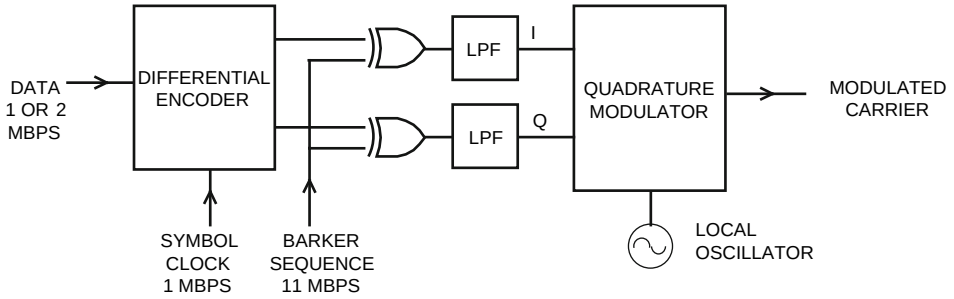


Figure 5.2-9 DSSS modulation.

circuits correlate the Barker sequence with the input signal and output an analog signal that represents the degree of correlation. The following differential decoder performs the opposite operation of the differential encoder described above and outputs the 1 or 2 Mbps data.

The process of despreading the input signal by correlating it with the stored spreading sequence requires synchronization of the receiver with transmitter timing and frequency. To facilitate this, the transmitted frame starts with a synchronization field (SYNC), shown at the beginning of the physical layer protocol data unit in Figure 5.2-11. Then a start frame delimiter (SFD) marks out the commencement of the following information bearing fields. All bits in the indicated preamble are transmitted at a rate of 1 Mbps, no matter what the subsequent data rate will be. The signal field specifies the data rate of the following fields in the frame so that the receiver can adjust itself accordingly. The next field, SERVICE, contains all zeros for devices that are only compliant with the basic version of 802.11, but some of its bits are used in devices conforming with updated versions. The value of the length field is the length, in microseconds, required to transmit the data-carrying field labeled MPDU (MAC protocol data unit). An error check field, labeled CRC, protects the integrity of the SIGNAL, SERVICE, and LENGTH fields. The last field MPDU (MAC protocol data unit) is the data passed down from the MAC to be sent by the physical layer, or to be passed up to the MAC after reception. All bits in the transmitted frame are pseudo-randomly scrambled to ensure even power distribution over the spectrum. Data is returned to its original form by descrambling in the receiver.

802.11b

The "b" supplement to the original 802.11 specification supports a higher rate physical layer for the 2.4 GHz band. It is this 802.11b version that provided the impetus for Wi-Fi proliferation. With it, data rates of 5.5 Mbps and 11 Mbps are enabled, while retaining downward compatibility with the original 1 and 2 Mbps rates. The slower rates may be used not only for compatibility with devices that aren't capable of the extended rates, but also for fall back when interference or range conditions don't provide the required signal-to-noise ratio for communication using the higher rates.

As previously stated, the increased data rates provided for in 802.11b do not entail a larger channel bandwidth. Also, the narrow-band interference rejection, or jammer resisting qualities of direct sequence spread-spectrum are retained. The classical definition of processing gain for DSSS as being the chip rate divided by the data bandwidth doesn't apply here. In fact, the processing gain requirement that for years was part of the FCC Rules paragraph 15.247 definition of direct sequence spread-spectrum was deleted in an update from August 2002, and at the same time reference to DSSS was replaced by "digital modulation."

The mandatory high-rate modulation method of 802.11b is called complementary code keying (CCK). An optional mode called packet binary convolutional

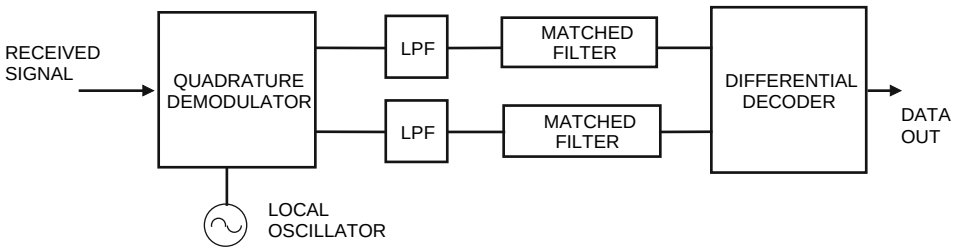


Figure 5.2-10 DSS reception.

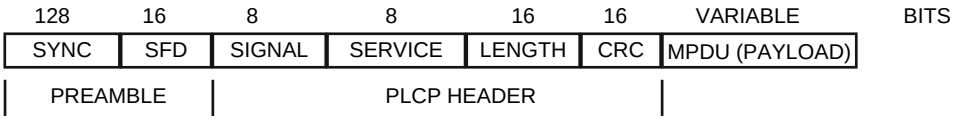


Figure 5.2-11 DSSS frame format.

coding (PBCC) is also described in the specification. Although there are similarities in concept, the two modes differ in implementation and performance. First the general principle of high-rate DSSS is presented below, applying to both CCK and PBCC, then the details of CCK are given.

As in the original 802.11, a pseudo-random noise sequence at the rate of 11 Mcps is the basis of high-rate transmission in 802.11b. It is this 11 Mcps modulation that gives the 22 MHz null-to-null bandwidth. However, in contrast to the original specification, the symbol rate when sending data at 5.5 or 11 Mbps is 1.375 Msps. Eight chips per symbol are transmitted instead of eleven chips per symbol as when sending at 1 or 2 Mbps. In "standard" DSSS as used in 802.11, the modulation, BPSK or QPSK, is applied to the group of eleven chips constituting a symbol. The series of eleven chips in the symbol is always the same (the Barker sequence previously defined). In contrast, high-rate DSSS uses a different 8-chip sequence in each symbol, depending on the sequence of data bits that is applied to each symbol. Quadrature modulation is used, and each chip has an I value and a Q value which represent a complex number having a normalized amplitude of one and some angle, α, where α = arctangent (Q/I). α can assume one of four values divided equally around 360 degrees. Since each complex bit has four possible values, there are a total of $4^8 = 65536$ possible 8-bit complex words. For the 11 Mbps data rate, 256 out of these possibilities are actually used—which one being determined by the sequence of 8 data bits applied to a particular symbol. Only 16-chip sequences are needed for the 5.5 Mbps rate, determined by four data bits per symbol. The high-rate algorithm describes the manner in which the 256 code words, or 16 code words, are chosen from the 65536 possibilities. The chosen 256 or 16 complex words have the very desirable property that when correlation detectors are used on the I and Q lines of the received signal, downconverted to baseband, the original 8-bit (11 Mbps rate) or 4-bit (5.5 Mbps rate) sequence can be decoded correctly with high probability even when reception is accompanied by noise and other types of channel distortion.

The concept of CCK modulation and demodulation is shown in Figures 5.2-12 and 5.2-13. It's explained below in reference to a data rate of 11 Mbps. The multiplexer of Figure 5.2-12 takes a block of eight serial data bits, entering at 11 Mbps, and outputs them in parallel, with updates at the symbol rate of 1.375 MHz. The six latest data bits determine 1 out of 64 (2^6) complex code words. Each code word is a sequence of eight complex chips, having phase angles α_1 through α_8 and a magnitude of unity. The first two data bits, d_0 and d_1 determine an angle, α'_8 which, in the code rotator (see Figure 5.2-12), rotates the whole code word relative to α_8 of the previous code word. This angle of rotation becomes the absolute angle α_8 of the present code word. The normalized I and Q outputs of the code rotator, which after filtering are input to a quadrature modulator for up-conversion to the carrier (or intermediate) frequency, are:

$$I_i = \cos(\alpha_i), \quad Q_i = \sin(\alpha_i) \quad i = 1 \ldots 8.$$

Figure 5.2-13 is a summary of the development of code words a for 11 Mbps rate CCK modulation. High rate modulation is applied only to the payload —MPDU in Figure 5.2-11. The code word described in Figure 5.2-13 is used as shown for the first symbol and then every other symbol of the payload. However, it is

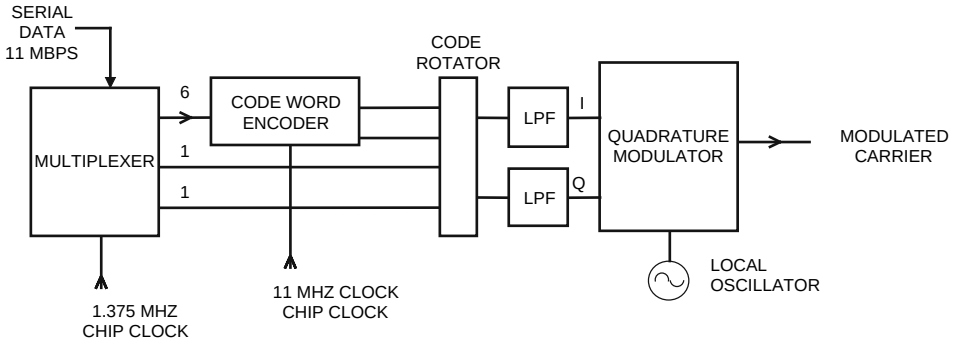


Figure 5.2-12 High-rate modulator—11 Mbps.

Data symbol: $d_0\ d_1\ d_2\ d_3\ d_4\ d_5\ d_6\ d_7$

Phase Table		
d_i	d_{i+1}	φ
0	0	0^0
1	0	180^0
0	1	90^0
1	1	-90^0

Phase $(d_0,d_1) = \varphi_1$
Phase $(d_2,d_3) = \varphi_2$
Phase $(d_4,d_5) = \varphi_3$
Phase $(d_6,d_7) = \varphi_4$

$\alpha_1 = \varphi_1 + \varphi_2 + \varphi_3 + \varphi_4$
$\alpha_2 = \varphi_1 + \varphi_3 + \varphi_4$
$\alpha_3 = \varphi_1 + \varphi_2 + \varphi_4$
$\alpha_4 = \varphi_1 + \varphi_4 + 180^0$
$\alpha_5 = \varphi_1 + \varphi_2 + \varphi_3$
$\alpha_6 = \varphi_1 + \varphi_3$
$\alpha_7 = \varphi_1 + \varphi_2 + 180^0$
$\alpha_8 = \varphi_1$

$I_i = \cos(\alpha_i)$
$Q_i = \sin(\alpha_i)$
$i = 1 \ldots 8$

Figure 5.2-13 Derivation of code word.

modified by adding 180° to each element of the code word of the second symbol, fourth symbol, and so on.

The development of the symbol code word or chip sequence may be clarified by an example worked out per Figure 5.2-13. Let's say the 8-bit data sequence for a symbol is $\boldsymbol{d} = d_0 \ldots d_7 = 1\ 0\ 1\ 0\ 1\ 1\ 0\ 1$. From the phase table of Figure 5.2-13 we find the angles φ: $\varphi_1 = 180°$, $\varphi_2 = 180°$, $\varphi_3 = -90°$, $\varphi_4 = 90°$. Now summing up these values to get the angle α_i of each complex chip, then taking the cosine and sine to get I_i and Q_i, we summarize the result in the following table:

i	1	2	3	4	5	6	7	8
α	0	180	90	90	−90	90	180	180
I	1	−1	0	0	0	0	−1	−1
Q	0	0	1	1	−1	1	0	0

The code words for 5.5 Mbps rate CCK modulation are a subset of those for 11 Mbps CCK. In this case, there are four data bits per symbol which determine a total of 16 complex chip sequences. Four 8-element code words (complex chip sequences) are determined using the last two data bits of the symbol, d_2 and d_3. The arguments (angles) of these code words are shown in Table 5.2-1. Bits d_0 and d_1 are used to rotate the code words relative to the preceding code word as in 11 Mbps modulation and shown in the phase table of Figure 5.2-13. Code words are modified by 180° every other symbol, as in 11 Mbps modulation.

Table 5.2-1 5. 5 mbps CCK decoding.

d_3,d_2	α_1	α_2	α_3	α_4	α_5	α_6	α_7	α_8
00	90°	0°	90°	180°	90°	0°	−90°	0°
10	−90°	180°	−90°	0°	90°	0°	−90°	0°
01	−90°	0°	−90°	180°	−90°	0°	90°	0°
11	90°	180°	90°	0°	−90°	0°	90°	0°

The concept of CCK decoding for receiving high rate data is shown in Figure 5.2-14. For the 11 Mbps data rate, a correlation bank decides which of the 64 possible codes best fits each received 8-bit symbol. It also finds the rotation angle of the whole code relative to the previous symbol (one of four values). There are a total of 256 (64 × 4) possibilities and the chosen one is output as serial data. At the 5.5 Mbps rate there are four code words to choose from and after code rotation a total of 16 choices from which to decide on the output data.

To maintain compatibility with earlier non high-rate systems, the DSSS frame format shown in Figure 5.2-11 is retained in 802.11b. The 128-bit preamble and the header are transmitted at 1 Mbps while the payload MPDU can be sent at a high rate of 5.5 or 11 Mbps. The long and slow preamble reduces the throughput and cancels some of the advantage of the high data rates. 802.11b defines an optional short preamble and header which differ from the standard frame by sending a preamble with only 72 bits and transmitting the header at 2 Mbps, for a total overhead of 96 μsec instead of 192 μsec for the long preamble and header. Devices using this option can only communicate with other stations having the same capability.

Use of higher data rates entails some loss of sensitivity and hence range. The minimum specified sensitivity at

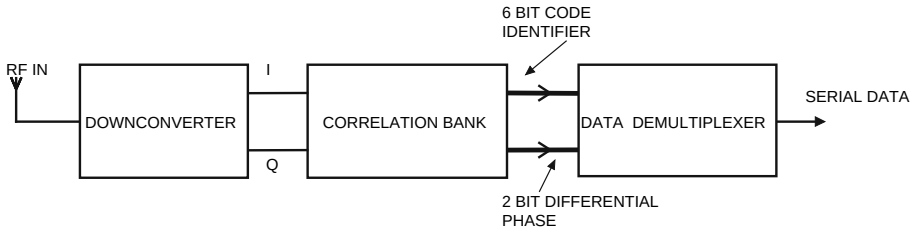


Figure 5.2-14 CCK decoding.

the 11 Mbps rate is –76 dBm for a frame-error rate of 8% when sending a payload of 1024 bytes, as compared to a sensitivity of –80 dBm for the same frame-error rate and payload length at a data rate of 2 Mbps.

802.11a and OFDM

In the search for ways to communicate at even higher data rates than those applied in 802.11b, a completely different modulation scheme, OFDM (orthogonal frequency division multiplexing) was adopted for 802.11a. It is not DSSS yet it has a channel bandwidth similar to the DSSS systems already discussed. The 802.11a supplement is defined for channel frequencies between 5.2 and 5.85 GHz, obviously not compatible with 802.11b signals in the 2.4 GHz band. However, since the channel occupancy characteristics of its modulation are similar to that of DSSS Wi-Fi, the same system was adopted in IEEE 802. 11g for enabling the high data rates of 802.11a on the 2.4 GHz band, while allowing downward compatibility with transmissions conforming to 802.11b.

802.11a specifies data rates of 6, 9, 12, 18, 24, 36, 48, and 54 Mbit/s. As transmitted data rates go higher and higher, the problem of multipath interference becomes more severe. Reflections in an indoor environment can result in multipath delays on the order of 100 nsec but may be as long as 250 nsec, and a signal with a bit rate of 10 Mbps (period of 100 nsec) can be completely overlapped by its reflection. When there are several reflections, arriving at the receiver at different times, the signal may be mutilated beyond recognition. The OFDM transmission system goes a long way to solving the problem. It does this by sending the data partitioned into symbols whose length in time is several times the expected reflected path length time differences. The individual data bits in a symbol are all sent in parallel on separate subcarrier frequencies within the transmission channel. Thus, by sending many bits during the same time, each on a different frequency, the individual transmitted bit can be lengthened so that it won't be affected by the multipath phenomenon. Actually, the higher bit rates are accommodated by representing a group of data bits by the phase and amplitude of a particular transmitted carrier. A carrier modulated using quadrature phase shift keying (QPSK) can represent two data bits and 64-QAM (quadrature amplitude modulation) can present six data bits as a single data unit on a subcarrier.

Naturally, transmitting many subcarriers on a channel of given width brings up the problem of interference between those subcarriers. There will be no interference between them if all the subcarriers are orthogonal—that is, if the integral of any two different subcarriers over the symbol period is zero. It is easy to show that this condition exists if the frequency difference between adjacent subcarriers is the inverse of the symbol period.

In OFDM, the orthogonal subcarriers are generated mathematically using the inverse Fourier transform (IFT), or rather its discrete equivalent, the inverse discrete Fourier transform (IDFT). The IDFT may be expressed as:

$$x(n) = \frac{1}{N}\sum_{m=0}^{N-1} X(m)[\cos(2\pi mn/N) + j\cdot\sin(2\pi mn/N)]$$

$x(n)$ are complex sample values in the time domain, $n = 0\ldots N-1$, and $X(m)$ are the given complex values, representing magnitude and phase, for each frequency in the frequency domain. The IDFT expression indicates that the time domain signal is the sum of N harmonically related sine and cosine waves each of whose magnitude and phase is given by $X(m)$. We can relate the right side of the expression to absolute frequency by multiplying the arguments $2\pi mn/N$ by f_s/f_s to get

$$x(n) = \frac{1}{N}\sum_{m=0}^{N-1} X(m)[\cos(2\pi mf_1nt_s) + j\cdot\sin(2\pi mf_1nt_s)] \quad (5.2.1)$$

where f_1 is the fundamental subcarrier and the difference between adjacent subcarriers, and t_s is the sample time $1/f_s$. In 802.11a OFDM, the sampling frequency is 20 MHz and $N = 64$, so $f_1 = 312.5$ kHz. Symbol time is $Nt_s = 64/f_s = 3.2$ μsec.

In order to prevent intersymbol interference, 802.11a inserts a guard time of 0.8 μsec in front of each symbol, after the IDFT conversion. During this time, the last 0.8 μsec of the symbol is copied, so the guard time is also called a circular prefix. Thus, the extended symbol time that is transmitted is $3.2 + .8 = 4$ μsec. The guard time is deleted after reception and before reconstruction of the transmitted data.

Although the previous equation, where $N = 64$, indicates 64 possible subcarriers, only 48 are used to carry data, and four more for pilot signals to help the receiver phase lock to the transmitted carriers. The remaining carriers that are those at the outside of the occupied bandwidth, and the DC term ($m = 0$ in Eq. (5.2.1)), are null. It follows that there are 26 ((48 + 4)/2) carriers on each side of the nulled center frequency. Each channel width is 312.5 kHz, so the occupied channels have a total width of 16.5625 (53 × 312.5 kHz) MHz.

For accommodating a wide range of data rates, four modulation schemes are used—BPSK, QPSK, 16-QAM and 64-QAM, requiring 1, 2, 4, and 6 data bits per symbol, respectively. Forward error correction (FEC) coding is employed with OFDM, which entails adding code bits in each symbol. Three coding rates: 1/2, 2/3, and 3/4, indicate the ratio of data bits to the total number of bits per symbol for different degrees of coding

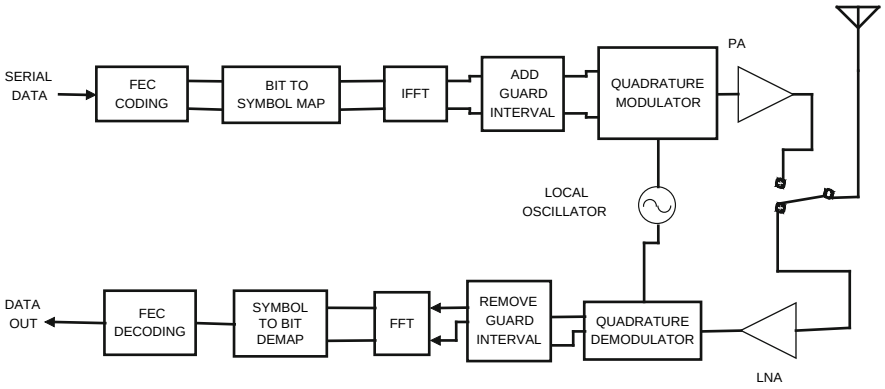


Figure 5.2-15 OFDM system block diagram.

performance. FEC permits reconstruction of the correct message in the receiver, even when one or more of the 48 data channels have selective interference that would otherwise result in a lost symbol. Symbol bits are interleaved so that even if adjacent subcarrier bits are demodulated with errors, the error correction procedure will still reproduce the correct symbol. A block diagram of the OFDM transmitter and receiver is shown in Figure 5.2-15. Blocks FFT and IFFT indicate the fast Fourier transform and its inverse instead of the mathematically equivalent (in terms of results) discrete Fourier transform and inverse discrete Fourier transform (IFDT) that we used above because it is much faster to implement. Table 5.2-2 lists the modulation type and coding rate used for each data rate, and the total number of bits per OFDM symbol, which includes data bits and code bits.

The available frequency channels in the 5 GHz band in accordance with FCC paragraphs 15.401–15.407 for unlicensed national information infrastructure (U-NII) devices are shown in Table 5.2-3. Channel allocations are 5 MHz apart and 20 MHz spacing is needed to prevent co-channel interference. Twelve simultaneous networks can coexist without mutual interference. Power limits are also shown in Table 5.2-4.

Extension of the data rates of 802.11b to those of 802.11a, but on the 2.4 GHz band is covered in supplement 802.11g. The OFDM physical layer defined for the 5 GHz band is applied essentially unchanged to 2.4 GHz. Equipment complying with 802.11g must also have the lower-rate features and the CCK modulation technique of 802.11b so that it will be downward compatible with existing Wi-Fi systems.

HIPERLAN/2

While 802.11b was designed for compliance with regulations in the European Union and most other regions of the world, 802.11a specifically refers to the regulations of the FCC and the Japanese MPT. ETSI (European Telecommunications Standards Institute) developed a high-speed wireless LAN specification, called HIPERLAN/2 (high performance local area network), which meets the European regulations and in many ways goes beyond the capabilities of 802.11a. HIPERLAN/2 defines a physical layer essentially identical to that of 802.11a, using coded OFDM and the same data rates up to 54 Mbps. However, its second layer software level is very different from the 802.11 MAC and the two systems are not compatible. Built-in features of HIPERLAN/2 that distinguish it from IEEE 802.11a are the following:

- *Quality of service* (QOS). Time division multiple access/time division duplex (TDMA/TDD) protocol permits multimedia communication.

Table 5.2-2 OFDM characteristics according to data rate.

Data rate Mbps	Modulation	Coding rate	Coded bits per subcarrier	Coded bits per OFDM symbol	Data bits per OFDM symbol
6	BPSK	1/2	1	48	24
9	BPSK	3/4	1	48	36
12	QPSK	1/2	2	96	48
18	QPSK	3/4	2	96	72
24	16-QAM	1/2	4	192	96
36	16-QAM	3/4	4	192	144
48	64-QAM	2/3	6	288	192
54	64-QAM	3/4	6	288	216

Table 5.2-3 Channel allocations and maximum power for 802.11a in United States.

Band	Operation channel numbers	Channel center frequencies (MHz)	Maximum power with up to 6 dBi antenna gain (mW)
U-NII lower band (5.15–5.25 GHz)	36 40 44 48	5180 5200 5220 5240	40
U-NII middle band (5.25–5.35 GHz)	52 56 60 64	5260 5280 5300 5320	200
U-NII upper band (5.725–5.825 GHz)	149 153 157 161	5745 5765 5785 5805	800

- *Dynamic frequency selection* (*DFS*). Network channels are selected and changed automatically to maintain communication reliability in the presence of interference and path disturbances.
- *Transmit power control (TPC)*. Transmission power is automatically regulated to reduce interference to other frequency band users and reduce average power supply consumption.
- *High data security*. Strong authentication and encryption procedures.

All of the above features of HIPERLAN/2 are being dealt with by IEEE task groups for implementation in 802.11. Specifically, the features of DFS and TPC are necessary for conformance of 802.11a to European Union regulations.

Frequency channels and power levels of HIPERLAN/2 are shown in Table 5.2-4.

Table 5.2-4 HIPERLAN/2 frequency channels and power levels (ETSI TS 101 475 V1.3.1 (2001–12)).

Center frequency (MHz)	Radiated power (mean EIRP) (dBm)
Every 20 MHz from 5180 to 5320	23
Every 20 MHz from 5500 to 5680	30
5700	23

5.2.2 Bluetooth

There are two sources of the Bluetooth specification. One is the Bluetooth Special Interest Group (SIG). The current version at this writing is Version 1.1. It is arranged in two volumes—Core and Profiles. Volume 1, the core, describes the physical, or hardware radio characteristics of Bluetooth, as well as low-level software or firmware which serves as an interface between the radio and higher level specific user software. The profiles in Volume 2 detail protocols and procedures for several widely used applications. The other Bluetooth source specification is IEEE 802.15.1. It is basically a rewriting of the SIG core specification, made to fit the format of IEEE communications specifications in general.

Bluetooth is an example of a wireless personal area network (WPAN), as opposed to a wireless local area network (WLAN). It's based on the creation of *ad hoc*, or temporary, on-the-fly connections between digital devices associated with an individual person and located in the vicinity of around ten meters from him. Bluetooth devices in a network have the function of a master or a slave, and all communication is between a master and one or more slaves, never directly between slaves. The basic Bluetooth network is called a piconet. It has one master and from one to seven slaves. A scatternet is an interrelated network of piconets where any member of a piconet may also belong to an adjacent piconet. Thus, conceptually, a Bluetooth network is infinitely expandable. Figure 5.2-16 shows a scatternet made up of three piconets. In it, a slave in one piconet is a master in another. A device may be a master in one piconet only.

The basic RF communication characteristics of Bluetooth are shown in Table 5.2-5.

A block diagram of a Bluetooth transceiver is shown in Figure 5.2-17. It's divided into three basic parts: RF, baseband, and application software. A Bluetooth chip set will usually include the RF and baseband parts, with the application software being contained in the system's computer or controller. The user data stream originates and terminates in the application software. The baseband section manipulates the data and forms frames or data bursts for transmission. It also controls the frequency synthesizer according to the Bluetooth frequency-hopping protocol. The blocks in Figure 5.2-17 are general and various transmitter and receiver configurations are adopted by different manufacturers. The Gaussian low-pass filter block before the modulator, for example, may be implemented digitally as part of a complex signal *I*/*Q* modulation unit or it may be a discrete element filter whose output is applied to the frequency control line of a VCO. Similarly, the receiver may be one of several types. If a superheterodyne configuration is chosen, the filter at the output of the downconverter will be

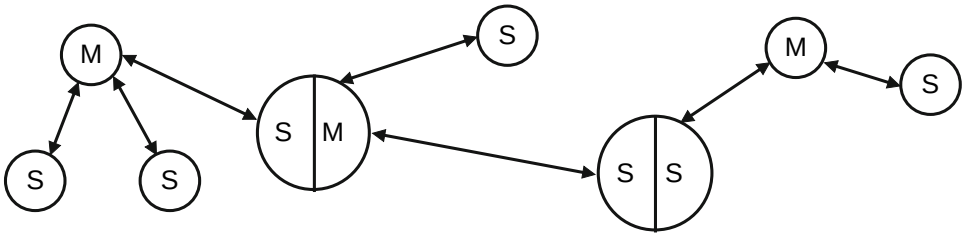


Figure 5.2-16 Bluetooth scatternet.

a bandpass type. A direct conversion receiver will use low pass filters in complex I and Q outputs of the down-converter. While different manufacturers employ a variety of methods to implement the Bluetooth radio, all must comply with the same strictly defined Bluetooth specification, and therefore the actual configuration used in a particular chipset should be of little concern to the end user.

The Bluetooth protocol has a fixed-time slot of 625 microseconds, which is the inverse of the hop rate given in Table 5.2-5. A transmission burst may occur within a duration of one, three, or five consecutive slots on one hop channel. As mentioned, transmissions are always between the piconet master and a slave, or several slaves in the case of a broadcast, or point-to-multipoint transmission. All slaves in the piconet have an internal timer synchronized to the master device timer, and the state of this timer determines the transmission hop frequency of the master and that of the response of a designated slave. Figure 5.2-18 shows a sequence of transmissions between a master and two slaves. Slots are numbered according to the state, or phase, of the master clock, which is copied to each slave when it joins the piconet. Note that master transmissions take place during even numbered clock phases and slave transmissions during odd numbered phases. Transmission frequency depends on the clock phase, and if a device makes a three or five slot transmission (slave two in the diagram), the intermediate frequencies that would have been used if only single slots were transmitted are omitted (f_4 and f_5 in this case). Note that transmissions do not take up a whole slot. Typically, a single-slot transmission burst lasts 366 microseconds, leaving 259 microseconds for changing the frequency of the synthesizer, phase locked loop settling time, and for switching the transceiver between transmit and receive modes.

Table 5.2-5 Bluetooth technical parameters

Characteristic	Value	Comment
Frequency band	2.4 to 2.483 GHz	May differ in some countries
Frequency hopping spread spectrum (FHSS)	79 1-MHz channels from 2402 to 2480 MHz	May differ in some countries
Hop rate	1600 hops per second	
Channel bandwidth	1 MHz	20 dB down at edges
Modulation	Gaussian frequency shift keying (GFSK)	
	Filter BT = 0.5	Gaussian Filter bandwidth = 500 kHz
	Nominal modulation index = 0.32	Nominal deviation = 160 kHz
Symbol rate	1 Mbps	
Transmitter maximum power		
Class 1	100 mW	Power control required
Class 2	2.5 mW	Must be at least 0.25 mW
Class 3	1 mW	No minimum specified
Receiver sensitivity	−70 dBm for BER = 0.1%	

There are two different types of wireless links associated with a Bluetooth connection. An asynchronous connectionless link (ACL) is used for packet data transfer while a synchronous connection oriented link (SCO) is primarily for voice. There are two major differences between the two link types. When an SCO link is established between a master and a slave, transmissions take place on dedicated slots with a constant interval between them. Also, unlike an ACL link, transmitted frames are not repeated in the case of an error in reception. Both of these conditions are necessary because voice is a continuous real-time process whose data rate cannot be randomly varied without affecting intelligibility. On the other hand, packet data transmission can use a handshaking protocol to regulate data accumulation and the instantaneous rate is not usually critical. Thus, for ACL links the master has considerable leeway in proportioning data transfer with the slaves in its network. An ARQ (automatic repeat request) protocol is always used, in addition to optional

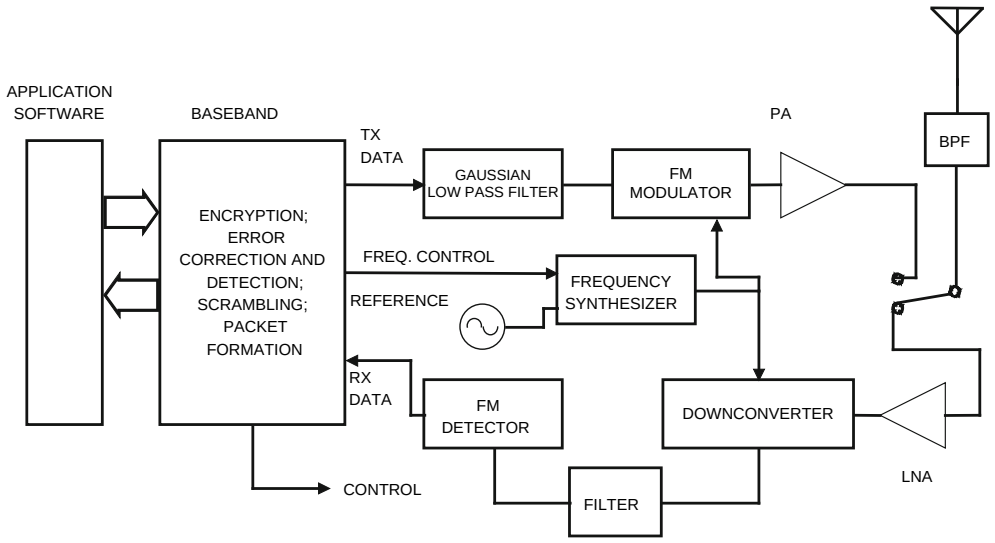


Figure 5.2-17 Bluetooth transceiver.

error correction, to ensure the highest reliability of the data transfer.

Bluetooth was conceived for employment in mobile and portable devices, which are more likely than not to be powered by batteries, so power consumption is an important issue. In addition to achieving low-power consumption due to relatively low transmitting power levels, Bluetooth incorporates power saving features in its communication protocol. Low average power is achieved by reducing the transmission duty cycle, and putting the device in a low-power standby mode for as long a period as possible relative to transmit and receive times while still maintaining the minimum data flow requirements. Bluetooth has three modes for achieving different degrees of power consumption during operation: sniff, hold, and park. Even in the normal active mode, some power saving can be achieved, as described below.

Active mode

During normal operation, a slave can transmit in a particular time slot only if it is specifically addressed by the master in the proceeding slot. As soon as it sees that its address is not contained in the header of the master's message, it can "go to sleep," or enter a low-power state until it's time for the next master transmission. The master also indicates the length of its transmission (one, three, or five slots) in its message header, so the slave can extend its sleep time during a multiple slot interval.

Sniff mode

In this mode, sleep time is increased because the slave knows in advance the time interval between slots during which the master may address the slave. If it's not addressed during the agreed slot, it returns to its low-power state for the same period and then wakes up and listens again. When it is addressed, the slave continues listening during subsequent master transmission slots as long as it is addressed, or for an agreed time-out period.

Hold mode

The master can put a slave in the hold mode when data transfer between them is being suspended for a given period of time. The slave is then free to enter a low-power state, or do something else, like participate in another piconet. It still maintains its membership in the original piconet, however. At the end of the agreed time interval, the slave resynchronizes with the traffic on the piconet and waits for instructions from the master.

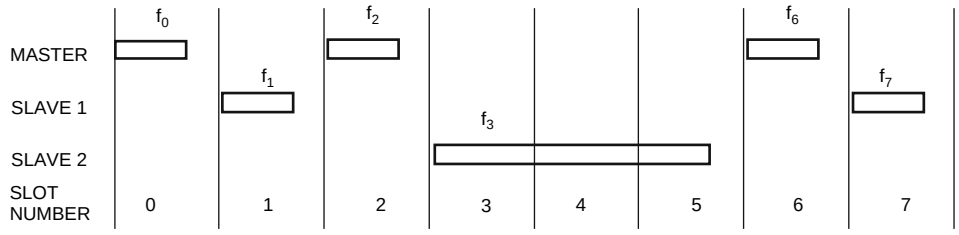


Figure 5.2-18 Bluetooth timing.

Park mode

Park has the greatest potential for power conservation, but as opposed to hold and sniff, it is not a directly addressable member of the piconet. While it is outside of direct calling, a slave in park mode can continue to be synchronized with the piconet and can rejoin it later, either on its own initiative or that of the master, in a manner that is faster than if it had to join the piconet from scratch. In addition to saving power, park mode can also be considered a way to virtually increase the network's capacity from eight devices to 255, or even more. When entering park mode, a slave gives up its active piconet address and receives an 8-bit parked member address. It goes into low-power mode but wakes up from time to time to listen to the traffic and maintain synchronization. The master sends beacon transmissions periodically to keep the network active. Broadcast transmissions to all parked devices can be used to invite any of them to rejoin the network. Parked units themselves can request re-association with the active network by way of messages sent during an access window that occurs a set time after what is called a "beacon instant." A polling technique is used to prevent collisions.

Packet format

In addition to the data that originates in the high-level application software, Bluetooth packets contain fields of bits that are created in the baseband hardware or firmware for the purpose of acquisition, addressing, and flow control. Packet bits are also subjected to data whitening (randomization), error-correction coding, and encryption as defined for each particular data type. Figure 5.2-19 shows the standard packet format.

The access code is used for synchronization, d-c level compensation, and identification. Each Bluetooth device has a unique address, and it is the address of the device acting as master that is used to identify transmitted packets as belonging to a specific piconet. A 64-bit synchronization word sandwiched between a four-bit header and four-bit trailer, which provide d-c compensation, is based on the master's address. This word has excellent correlation properties so when it is received by any of the piconet members it provides synchronization and positive identification that the packet of which it is a part belongs to their network. All message packets sent by members of the piconet use the same access code.

The header contains six fields with link control information. First, it has a three-bit active member address which identifies to which of the up to seven slaves a master's message is destined. An all zero address signifies a broadcast message to all slaves in the piconet. The next field has four bits that define the type of packet being sent. It specifies, for example, whether one, three, or five slots are occupied, and the level of error correction applied. The remaining fields involve flow control (handshaking), error detection and sequencing. Since the header has prime importance in the packet, it is endowed with forward-error correction having a redundancy of times three.

Following the header in the packet is the payload, which contains the actual application or control data being transferred between Bluetooth devices. The contents of the payload field depend on whether the link is an ACL or SCO. The payload of ACL links has a payload header field that specifies the number of data bytes and also has a handshaking bit for data-buffering control. A CRC (cyclic redundancy check) field is included for data integrity. As stated above, SCO links don't retransmit packets so they don't include a CRC. They don't need a header either because the SCO payload has a constant length.

The previous packet description covers packets used to transfer user data, but other types of packets exist. For example, the minimum length packet contains only the access code, without the four-bit trailer, for a total of 68 bits. It's used in the inquiry and paging procedures for initial frequency-hopping synchronization. There are also NULL and POLL packets that have an access code and header, but no payload. They're sent when slaves are being polled to maintain synchronization or confirm packet reception (in the case of NULL) in the piconet but there is no data to be transferred.

Error correction and encryption

The use of forward error correction (FEC) improves throughput on noisy channels because it reduces the number of bad packets that have to be retransmitted. In the case of SCO links that don't use retransmission, FEC can improve voice quality. However, error correction involves bit redundancy so using it on relatively noiseless links will decrease throughput. Therefore, the application decides whether to use FEC or not.

As already mentioned, there are various types of packets, and the packet type defines whether or not FEC

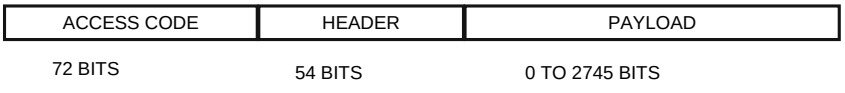


Figure 5.2-19 Bluetooth packet.

is used. The most redundant FEC method is always used in the packet header, and for the payload in one type of SCO packet. It simply repeats each bit three times, allowing the receiver to decide on the basis of majority rule what data bit to assign to each group of incoming bits.

The other FEC method, applied in certain type ACL and SCO packets, uses what's called a (15,10) shortened Hamming code. For every ten data bits, five parity bits are generated. Since out of every 15 transmitted bits only ten are retrieved, the data rate is only two-thirds what it would be without coding. This code can correct all single errors and detect all double errors in each 15-bit code word.

Wireless communication is susceptible to eavesdropping so Bluetooth incorporates optional security measures for authentication and encryption. Authentication is a procedure for verifying that received messages are actually from the party we expect them to be and not from an outsider who is inserting false messages. Encryption prevents an eavesdropper from understanding intercepted communications, since only the intended recipient can decipher them. Both authentication and implementation routines are implemented in the same way. They involve the creation of secret keys that are generated from the unique Bluetooth device address, a PIN (personal identification number) code, and a random number derived from a random or pseudo-random process in the Bluetooth unit. Random numbers and keys are changed frequently. The length of a key is a measure of the difficulty of cracking a code. Authentication in Bluetooth uses a 128-bit key, but the key size for encryption is variable and may range from 8 to 128 bits.

Inquiry and paging

A distinguishing feature of Bluetooth is its *ad hoc* protocol and connections are often required between devices that have no previous knowledge of their nature or address. Also, Bluetooth networks are highly volatile, in comparison to WLAN for example, and connections are made and dissolved with relative frequency. To make a new connection, the initiator—the master—must know the address of the new slave, and the slave has to synchronize its clock to the master's in order to align transmit and receive channel hop-timing and frequencies. The inquiry and paging procedures are used to create the connections between devices in the piconet.

By use of the inquiry procedure, a connection initiator creates a list of Bluetooth devices within range. Later, desired units can be summoned into the piconet of which the initiator is master by means of the paging routine.

As mentioned previously, the access code contains a synchronization word based on the address of the master. During inquiry, the access code is a general inquiry access code (GIAC) formed from a reserved address for this purpose. Dedicated inquiry access codes (DIAC) can also be used when the initiator is looking only for certain types of devices. Now a potential slave can lock on to the master, provided it is receiving during the master's transmission time and on the transmission frequency. To facilitate this match-up, the inquiry procedure uses a special frequency hop routine and timing. Only 32 frequency channels are used and the initiator transmits two burst hops per standard time slot instead of one. On the slot following the transmission inquiry bursts, the initiator listens for a response from a potential slave on two consecutive receive channels whose frequencies are dependent on the previously transmitted frequencies.

When a device is making itself available for an inquiring master, it remains tuned to a single frequency for a period of 1.28 seconds and at a defined interval and duration scans the channel for a transmission. At the end of the 1.28-second period, it changes to another channel frequency. Since the master is sending bursts over the whole inquiry frequency range at a fast rate—two bursts per 1250 microsecond interval—there's a high probability the scanning device will catch at least one of the transmissions while it remains on a single frequency. If that channel happens to be blocked by interference, then the slave will receive a transmission after one of its subsequent frequency changes. When the slave does hear a signal, it responds during the next slot with a special packet called FHS (frequency hop synchronization) in which is contained the slave's Bluetooth address and state of its internal clock register. The master does not respond but notes the slave's particulars and continues inquiries until it has listed the available devices in its range. The protocol has provisions for avoiding collisions from more than one scanning device that may have detected a master on the same frequency and at the same time.

The master makes the actual connection with a new device appearing in its inquiry list using the page routine. The paging procedure is quite similar to that of the inquiry. However, now the master knows the paged device's address and can use it to form the synchronization word in its access code. The designated slave does its page scan while expecting the access code derived from its own address. The hopping sequence is different during paging than during inquiry, but the master's transmission bursts and the slave's scanning routine are very similar.

A diagram of the page state transmissions is given in Figure 5.2-20. When the slave detects a transmission from the master (Step 1), it responds with a burst of access code based on its own Bluetooth address. The master then transmits the FHS, giving the slave the access code information (based on the master's address),

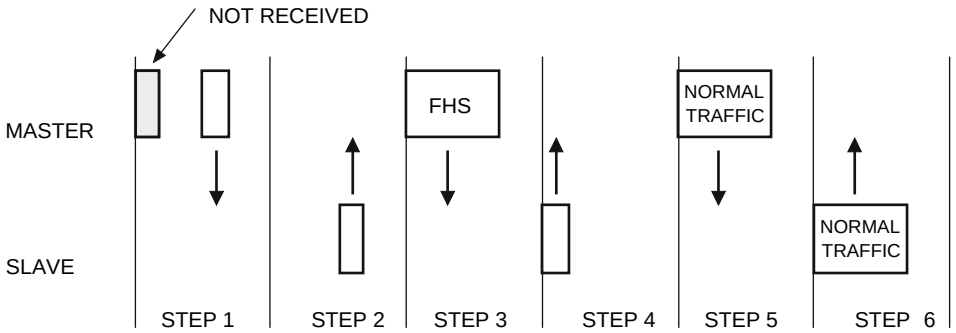


Figure 5.2-20 Paging transmissions.

timing and piconet active member address (between one and seven) needed to participate in the network. The slave acknowledges FHS receipt in Step 4. Steps 5 and 6 show the beginning of the network transmissions which use the normal 79 channel hopping-sequence based on the master's address and timing.

5.2.3 Zigbee

Zigbee is the name of a standards-based wireless network technology that addresses remote monitoring and control applications. Its promotion and development is being handled on two levels. A technical specification for the physical and data link layers, IEEE 802.15.4, was drawn up by a working group of the IEEE as a low data rate WPAN (wireless personal area network). An association of committed companies, the Zigbee Alliance, is defining the network, security, and application layers above the 802.15.4 physical and medium access control layers, and will deal with interoperability certification and testing.

The distinguishing features of Zigbee to which the IEEE standard addresses itself are

- Low data rates—throughput between 10 and 115.2 kbps
- Low power consumption—several months up to two years on standard primary batteries
- Network topology appropriate for multisensor monitoring and control applications
- Low complexity for low cost and ease of use
- Very high reliability and security

These will lend themselves to wide-scale use embedded in consumer electronics, home and building automation and security systems, industrial controls, PC peripherals, medical and industrial sensor applications, toys and games and similar applications. It's natural to compare Zigbee with the other WPAN standard, Bluetooth, and there will be some overlap in implementations. However, the two systems are quite different, as is evident from the comparison in Table 5.2-6.

Architecture

The basic architecture of Zigbee is similar to that of other IEEE standards, Wi-Fi and Bluetooth for example, a simplified representation of which is shown in Figure 5.2-21. On the bottom are the physical layers, showing two alternative options for the RF transceiver functions of the specification. Both of these options are never expected to exist in a single device, and indeed their transmission characteristics—frequencies, data rates, modulation system—are quite different. However, the embedded firmware and software layers above them will be essentially the same no matter what physical layer

Table 5.2-6 Comparison of Zigbee and Bluetooth

	Bluetooth	Zigbee
Transmission scheme	FHSS (frequency hopping spread spectrum)	DSSS (direct sequence spread spectrum)
Modulation	GFSK (gaussian frequency shift keying)	QPSK (quadrature phase shift keying) or BPSK (binary phase shift keying)
Frequency band	2.4 GHz	2.4 GHz, 915 MHz, 868 MHz
Raw data bit rate	1 Mbps	250 kbps, 40 kbps or 20 kbps (depends on frequency band)
Power output	Maximum 100 mW, 2.5 mW, or 1 mW, depending on class	Minimum capability 0.5 mW; maximum as allowed by local regulations
Minimum sensitivity	−70 dBm for 0.1% BER	−85 dBm (2.4GHz) or −92 dBm (915/868 MHz) for packet error rate <1%
Network topology	Master-Slave 8 active nodes	Star or Peer-Peer 255 active nodes

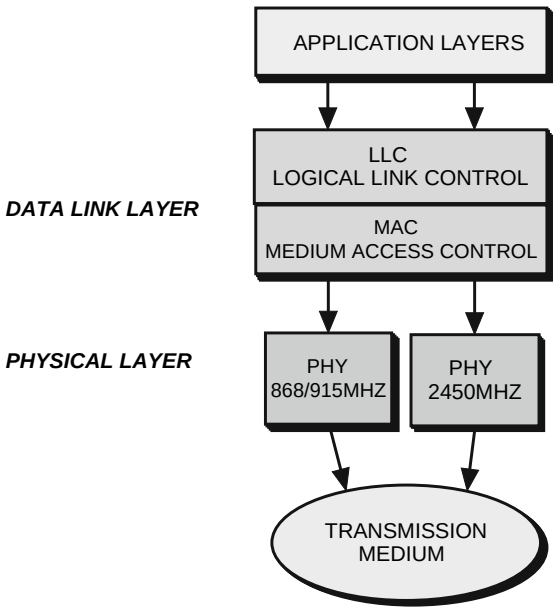


Figure 5.2-21 Zigbee architecture.

is applied. Just above the physical layers is the data link layer, consisting of two sublayers: medium access control, or MAC, and the logical link control, LLC. The MAC is responsible for management of the physical layer and among its functions are channel access, keeping track of slot times, and message delivery acknowledgement. The LLC is the interface between the MAC and physical layer and the upper-application software.

Application software is not a part of the IEEE 802.15.4 specification and it is expected that the Zigbee Alliance will prepare profiles, or programming guidelines and requirements for various functional classes in order to assure product interoperability and vendor independence. These profiles will define network formation, security, and application requirements while keeping in mind the basic Zigbee features of low power and high reliability.

Communication characteristics

In order to achieve high flexibility of adaptation to the range of applications envisioned for Zigbee, operation is being specified for three unlicensed bands—2.4 GHz, 915 MHz and 868 MHz; the latter two being included in the same physical layer. Those two bands are generally mutually exclusive, their use being determined by geographic location and regional regulations. The following 27 transmitting channels are defined:

Channel number	Center frequency range	Channel width
0	868.3 MHz	600 kHz
1 to 10	906 to 924 MHz	2 MHz
11 to 27	2405 to 2480	5 MHz

Data rates and modulation types for each of the bands are shown in Table 5.2-7.

In both physical layers, the modulation is DSSS (direct sequence spread spectrum). The spreading parameters are defined to meet communication authority regulations in the various regions as well as desired data rates. For example, the chip rate of 600 kbps on the 902–928 band allows the transmission to meet the FCC paragraph 15.247 requirement of minimum 500 kHz bandwidth at 6 dB down for digital modulation. However the chip rate, and with it the data rate, has to be reduced on Channel 0 in order to meet the confines of the 868 to 868.6 MHz channel allowed under ERC recommendation 70–03 and ETSI specification EN 300–220. On the 2400 to 2483.5 MHz band, the bit rate of 250 Kbps allows a throughput, after considering the overheads involved in packet transmissions, to attain 115.2 kbps, a rate used for some PC peripherals for example.

The spreading modulation used on the 2450 MHz physical layer has similarity in principle to that used on IEEE 802.11b (high-rate Wi-Fi) to increase the data bit rate without raising the chip rate, thereby achieving a desired carrier bandwidth. Sixteen different, almost orthogonal 32-bit long spreading sequences are available for transmission at 2 Mchips/second. Each consecutive sequence of four data bits determines which of the sixteen spreading sequences is sent. On reception, the receiver can identify the spreading sequence and thus decode the data bits. The modulation used, O-QPSK

Table 5.2-7 Data rates and modulation.

PHY (MHz)	Frequency band (MHz)	Spreading parameters		Data parameters		
		Chip rate (kcps)	Modulation	Bit rate (kbps)	Symbol rate (ksps)	Symbols
868/915	868–868.6 902–928	300 600	BPSK BPSK	20 40	20 40	Binary Binary
2450	2400–2483.5	2000	Offset-QPSK	250	62.5	16-ary orthogonal

(offset quadrature phase shift keying) with half-sine wave pulse shaping is essentially equivalent to a form of frequency shift keying, MSK (minimum shift keying). It is fairly easy to generate and has a relatively narrow bandwidth for the given chip rate. This latter feature allows a large number of nonoverlapping channels that can be used, with proper upper layer software, on the crowded 2.4 GHz band to avoid interference.

Other physical layer characteristics of Zigbee are output power and receiver sensitivity. The devices must be capable of radiating at least –3 dBm although output may be reduced to the minimum necessary in order to limit interference to other users. Maximum power is determined by the regulatory authorities. While much higher powers are allowed, it may not be practical to transmit over, say 10 dBm, because of absolute limits on spurious radiation and the general objective of low-cost and low-power consumption. Minimum receiver sensitivity for the 868/915 MHz physical layer is specified as –92 dBm and –85 dBm on 2.4 GHz. These limits are for a packet error rate of one percent.

Device types and topologies

Two device types, of different complexities, are defined. A full function device (FFD) will be able to implement the full protocol set and can act as a network coordinator. Devices capable of minimal protocol implementation are reduced function devices (RFD). Due to the distinction between device types, networks in which most members require only minimum functionality, such as switches and sensors, can be made significantly less costly and have lower power consumption than if all devices were constrained to have maximum capability.

Flexibility in network configuration is achieved through two topologies—star and peer-to-peer that are depicted in Figure 5.2-22. A network may have as many as 255 members, one of which is a PAN (personal area network) coordinator. The function of the PAN

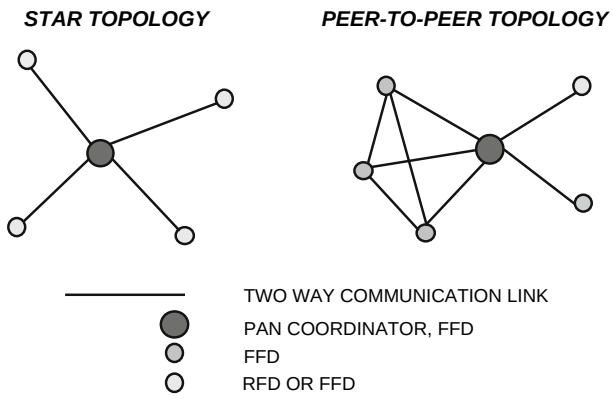


Figure 5.2-22 Network topologies.

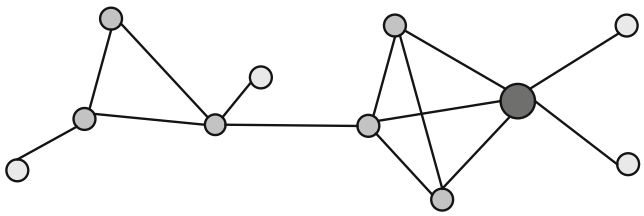


Figure 5.2-23 Cluster-tree topology.

coordinator, in addition to any specific application it may have, is to initiate, terminate, or route communication around the network. It also provides synchronization services. In a star network, each device communicates directly with the coordinator. The coordinator must be a FFD, and the others can be FFDs or RFDs. Relatively simple applications, like PC peripherals and toys, would typically use the star topology.

In the peer-to-peer topology, any device can communicate with any other device as long as it is in range. RFDs cannot participate, since an RFD can only communicate with a FFD. More complicated structures can be set up as a combination of peer-to-peer groups and star configurations. There is still just one PAN coordinator in the whole network. One example of such a structure is a cluster-tree network shown in Figure 5.2-23. In this arrangement devices on the network extremities may well be out of radio range of each other, but they can still communicate by relaying messages through the individual clusters.

Frame structure, collision avoidance, and reliability

Zigbee frame construction and channel access are similar to those of WLAN 802.11 (Wi-Fi) but are less complex. The transmitted packet has the basic construction shown in Figure 5.2-24. The purpose of the preamble is to permit acquisition of chip and symbol timing. The PHY header, which is signaled by a delimiter byte, notifies the baseband software in the receiver of the length of the subsequent data. The PSDU (PHY service data unit) is the message that has been passed down through the higher protocol layers. As shown, it can have a maximum of 127 bytes although monitoring and control applications will typically be much shorter. Included in the PSDU are information on the format of the message frame, a sequence number, address information, the data

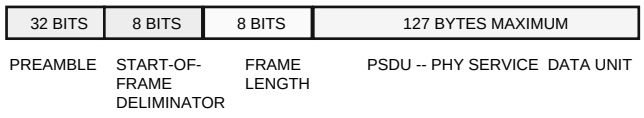


Figure 5.2-24 Transmission packet.

payload itself, and at the end, two bytes that serve as a frame check sequence. Reliability is assured since the receiver performs an independent calculation of this frame check sequence and compares it with the number received. If any bits have been changed by interference or noise, the numbers will not match. Only if a match occurs, the receiving side returns an acknowledgement to the originator of the message. Lacking an acknowledgement, the transmission will be repeated until it is successfully received.

In order to avoid two or more stations trying to transmit at the same time, a carrier sense multiple access with collision avoidance (CSMA-CA) routine is employed, similar to that used in Wi-Fi, IEEE 802.11. The Zigbee receiver monitors the channel and only if it is idle it may initiate a transmission. If the channel is occupied, the terminal must wait a random back off period before it can again attempt access.

Acknowledgement messages are sent without using the collision avoidance mechanism.

Zigbee applications

While the promoters of Zigbee aim to cover a very large market for those applications that require relatively low data rates, there will remain applications for which the compromises inherent in a general specification are not acceptable, and producers will continue to develop devices with proprietary specifications and characteristics. However, the open specification and a recognized certification of conformity are an advantage in many situations. For example, a home burglar alarm system would accept wireless sensors produced by different manufacturers, which will facilitate future expansion or allow installers to add sensors of types not available from the original system manufacturer. Use of devices approved according to a recognized standard gives the consumer some security against obsolescence.

Although Zigbee claims to be appropriate for most control applications, it will not fit all of them, and will not necessarily take advantage of all the possibilities of the unlicensed device regulations. Its declared maximum range of some 50 to 75 meters will fall short of the requirements of many systems. Given the meager maximum power allowed, greater range means reduced bandwidth and reduced data rate. In fact, a great many of the applications envisaged by Zigbee can get by very well with data rates of hundreds or a few thousand bits per second, and by matching receiver sensitivity to these rates, ranges of hundreds of meters can be achieved.

One partial answer to the range question is the deployment of the Zigbee network in a cluster-tree configuration, as previously described. Adjacent nodes serve as repeaters so that large areas can be covered, as long as the greatest distance between any two directly communicating nodes does not exceed Zigbee's basic range capability. For example, in a multi-floor building, sensors on the top floor can send alarms to the control box in the basement by passing messages through sensors located on every floor and operating as relay stations.

No doubt that there will be competition between Bluetooth and Zigbee for use in certain applications, but the overall deployment and the reliability of wireless control systems will increase. The proportion of wireless security and automation systems will increase because the new standard will provide a significant boost in reliability, security, and convenience, as compared to most present solutions.

5.2.4 Conflict and compatibility

With the steep rise of Bluetooth product sales and the already large and growing use of wireless local area networks, there is considerable concern about mutual interference between Bluetooth-enabled and Wi-Fi devices. Both occupy the 2.4 to 2.4835 GHz unlicensed band and use wideband spread-spectrum modulating techniques. They will most likely be operating concurrently in the same environments, particularly office/commercial but also in the home.

Interference can occur when a terminal of one network transmits on or near the receiving frequency of a terminal in another collocated network with enough power to cause an error in the data of the desired received signal. Although they operate on the same frequency band, the nature of Bluetooth and Wi-Fi signals are very different. Bluetooth has a narrowband transmission of 1 MHz bandwidth which hops around pseudo-randomly over an 80 MHz band while Wi-Fi (using DSSS) has a broad, approximately 20 MHz, bandwidth that is constant in some region of the band. The interference phenomenon is apparent in Figure 5.2-25. Whenever there is a frequency and time coincidence of the transmission of one system and reception of the other, it's possible for an error to occur. Whether it does or not depends on the relative signal strengths of the desired and undesired signals. These in turn depend on the radiated power outputs of the transmitters and the distance between them and the receiver. When two terminals are very close (on the order of centimeters), interference may occur even when the transmitting frequency is outside the bandwidth of the affected receiver.

Bluetooth and Wi-Fi systems are not synchronous and interference between them has to be quantified statistically. We talk about the probability of a packet error of one system caused by the other system. The consequence of a packet error is that the packet will have to be retransmitted once or more until it is correctly received,

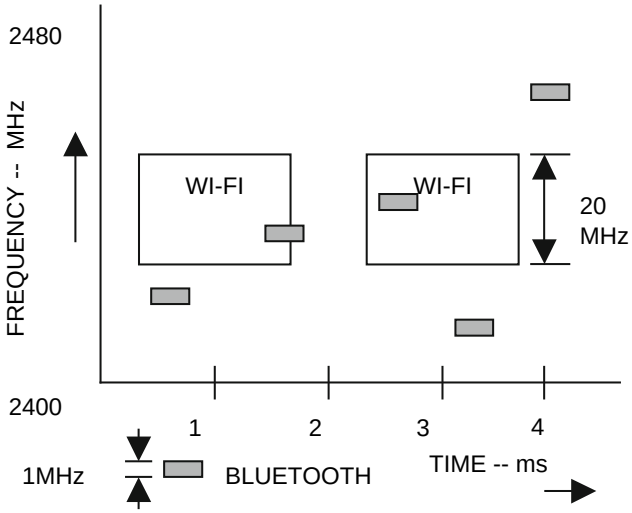


Figure 5.2-25 Wi-Fi and Bluetooth spectrum occupation.

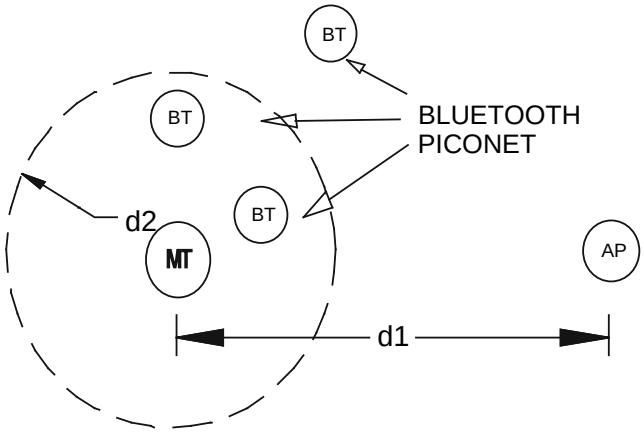


Figure 5.2-26 Importance of relative terminal location.

which causes a delay in message throughput. Voice transmissions generally don't allow packet retransmission because throughput cannot be delayed, so interference results in a decrease in message quality.

Following are parameters that affect interference between Bluetooth and Wi-Fi:

- *Frequency and time overlap.* A collision occurs when the interferer transmits at the same time as the desired transmitter and is strong enough to cause a bit or symbol error in the received packet.
- *Packet length.* The longer the packet length of the Wi-Fi system, relative to a constant packet length and hop rate of Bluetooth, the longer the victim may be exposed to interference from one or more collisions and the greater the probability of a packet error.
- *Bit rate.* Generally, the higher the bit rate, the lower the receiver sensitivity and therefore the more susceptible the victim will be to packet error for given desired and interfering signal strengths. On the other hand, higher bit rates usually result in reduced packet length, with the opposite effect.
- *Use factor.* Obviously, the more often the interferer transmits, the higher the probability of packet error. When both communicating terminals of the interferer are in the interfering vicinity of the victim the use factor is higher than if the terminals are further apart and one of them does not have adequate strength to interfere with the victim.
- *Relative distances and powers.* The received power depends on the power of the transmitter and its distance. Generally, Wi-Fi systems use more power than Bluetooth, typically 20 mW compared to 1 mW. Bluetooth Class 1 systems may transmit up to 100 mW, but their output is controlled to have only enough power to give a required signal level at the receiving terminal.
- Signal-to-interference ratio of the victim receiver, SIR, for a specified symbol or frame error ratio.
- Type of modulation, and whether error-correction coding is used.

A general configuration for the location of Wi-Fi and interfering Bluetooth terminals is given in Figure 5.2-26. In this discussion, only transmissions from the access point to the mobile terminal are considered. We can get an idea of the vicinity around the Wi-Fi mobile terminal in which operating Bluetooth terminals will affect transmissions from the access point to the mobile terminal by examining the following parameters:

CI_{cc}, CI_{ac}—Ratio of signal carrier power to co-channel or adjacent channel interfering power for a given bit or packet error rate (probability).

P_{WF}, P_{BT}—Wi-Fi and Bluetooth radiated power outputs.

$PL = Kd^r$—Path loss which is a function of distance d between transmitting and receiving terminals, and the propagation exponent r. K is a constant.

$d1$—Distance between Wi-Fi mobile terminal and access point.

$d2$—Radius of area around mobile terminal within which an interfering Bluetooth transmitter signal will increase the Wi-Fi bit error rate above a certain threshold.

PR_{WF}, PR_{BT}—Received powers from the access point and from the Bluetooth interfering transmitters.

$d2$ as a function of $d1$ is found as follows, using power in dBm:

1. $PR_{WF} = P_{WF} - 10\log(Kd1^r)$;
   $PR_{BT} = P_{BT} - 10\log(Kd2^r)$

2. $CI_{cc} = PR_{WF} - PR_{BT} = P_{WF} - 10\log(Kd1^r) - P_{BT} + 10\log(Kd2^r)$

3. $(CI_{cc} - P_{WF} + P_{BT})/10r = \log(d2/d1)$

4. $d2 = d1 \cdot 10^{(CI_{CC} - P_{WF} + P_{BT})/(10 \cdot r)}$ (5.2.1)

As an example, the interfering area radius $d2$ is now calculated from equation (5.2.1) using the following system parameters:

$CI_{cc} = 10$ dB, $P_{WF} = 13$ dBm, $P_{BT} = 0$ dBm,
$r = 2$ (free space)
Solving equation (5.2.1): $d2 = d1 \times .71$

In this case, if a Wi-Fi terminal is located 15 meters from an access point, for example, all active Bluetooth devices within a distance of 10.6 meters from it have the potential of interfering. Only co-channel interference is considered. Adjacent channel interference, if significant, would increase packet error probability because many more Bluetooth hop channels would cause symbol errors. However, the adjacent channel CI_{ac} is on the order of 45 dB lower than CI_{cc} and would be noticed only when Bluetooth is several centimeters away from the Wi-Fi terminal.

The effect of an environment where path loss is greater than in free space can be seen by using an exponent $r = 3$. For the same Wi-Fi range of 15 meters, the radius of Bluetooth interference becomes 11.9 meters.

While equation (5.2.1) does give a useful insight into the range where Bluetooth devices are liable to deteriorate Wi-Fi performance, its development did involve simplifications. It considered that the signal-to-interference ratio that causes the error probability to exceed a threshold is constant for all wanted signal levels, which isn't necessarily so. It also implies a step relationship between signal-to-interference ratio and performance degradation, whereas the effect of changing interference level is continuous. The propagation law used in the development is also an approximation.

Methods for improving Bluetooth and Wi-Fi coexistence

By dynamically modifying one or more system operating parameters according to detected interference levels, coexistence between Bluetooth and Wi-Fi can be improved. Some of these methods are discussed below.

Power control

Limiting transmitter power to the maximum required for a satisfactory level of performance will reduce interference to collocated networks. Power control is mandatory for Class 1 Bluetooth systems, where maximum power is 100 mW. The effect of the power on the interference radius is evident in equation (5.2.1). For example, in a Bluetooth piconet established between devices located over a spread of distances from the master, the master will use only the power level needed to communicate with each of the slaves in the network. Lack of power control would mean that all devices would communicate at maximum power and the collocated Wi-Fi system would be exposed to a high rate of interfering Bluetooth packets.

Adaptive frequency hopping

Wi-Fi and Bluetooth share approximately 25 percent of the total Bluetooth hop-span of 80 MHz. Probably the most effective way to avoid interference between the two systems is to restrict Bluetooth hopping to the frequency range not used by Wi-Fi. When there is no coordination or cooperation between collocated networks, the Bluetooth piconet master would have to sense the presence of Wi-Fi transmissions and modify the frequency-hopping scheme of the network accordingly. A serious obstacle to this method was lifted by a change to the FCC regulations governing spread-spectrum transmissions in the 2.4 GHz band. Previously, frequency hopping devices were committed to hopping over at least 75 pseudo-randomly selected hop channels. In August 2002, paragraph 15.247, according to which Bluetooth and Wi-Fi devices are regulated, was changed to allow a minimum of 15 nonoverlapping channels in the 2400 to 2483 MHz band. In addition, the regulation allows employing intelligent hopping techniques, when less than 75 hopping frequencies are used, to avoid interference with other transmissions, and also suppression of transmission on an occupied channel provided that there are a minimum of 15 hops. The Bluetooth specification is due to be modified to take advantage of the adaptive frequency hopping method of avoiding interference.

There are situations where adaptive frequency hopping may not be effective or may have a negative effect. When two or more adjacent Wi-Fi networks are operating concurrently, they will utilize different 22 MHz sections of the 2.4 GHz band—three nonoverlapping Wi-Fi channels are possible. In this case, Bluetooth may not be able to avoid collisions while using a minimum of 15 hop frequencies. In addition, if there are several Bluetooth piconets in the same area, collisions among themselves will be much more frequent than when the full 79 channel hopping sequences are used.

Packet fragmentation

The two interference-avoiding methods described above are applicable primarily for action by the Bluetooth network. One method that the Wi-Fi network can employ to improve throughput is packet fragmentation. By fragmenting data packets and sending more, but shorter transmission frames, each transmission will have a lower probability of collision with a Bluetooth packet. Although reducing frame size increases the percentage of overhead bits in the transmission, when interference is heavy the overall effect may be higher throughput than if fragmentation was not used. Increasing bit rate for a constant packet length will also result in a shorter transmitted frame and less exposure to interference.

The methods mentioned above for reducing interference presume no coordination between the two different types of collocated wireless networks. However, devices are now being produced, in laptop and notebook computers for example, that include both Wi-Fi and Bluetooth, sometimes even in the same chipset. In this case collaboration is possible in the device software to prevent inter-network collisions.

5.2.5 Ultra-wideband technology

Ultra-wideband (UWB) technology is based on transmission of very narrow electromagnetic pulses at a low repetition rate. The result is a radio spectrum that is spread over a very wide bandwidth—much wider than the bandwidth used in the spread-spectrum systems previously discussed. Ultra-wideband transmissions are virtually undetectable by ordinary radio receivers and therefore can exist concurrently with existing wireless communications without demanding additional spectrum or exclusive frequency bands.

- These are some of the advantages cited for ultra-wideband technology:
- Very low spectral density—Very low probability of interference with other radio signals over its wide bandwidth,
- High immunity to interference from other radio systems,
- Low probability of interception/detection by other than the desired communication link terminals,
- High multipath immunity,
- Many high data rate ultra-wideband channels can operate concurrently,
- Fine range-resolution capability,
- Relatively simple, low-cost construction, based on nearly all digital architectures.

Transmission and reception methods are unique, and are described briefly below.

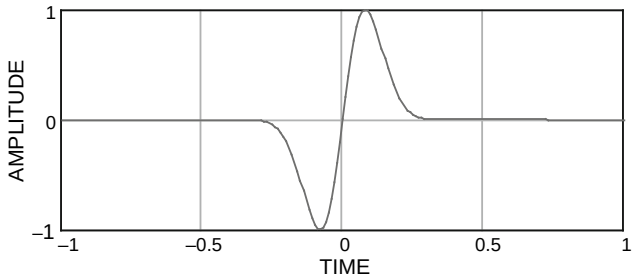


Figure 5.2-27 UWB monopulse.

Differing from conventional radio communication systems, which use up conversion and down conversion to pass information signals between baseband and bandpass frequency channels where wireless propagation occurs, UWB signal generation and detection use baseband techniques. An example of a UWB "carrier" is a Gaussian monopulse, shown in Figure 5.2-27. Its power spectrum is shown in Figure 5.2-28. If the time scale in Figure 5.2-27 is in nanoseconds, then the width of the pulse is 0.5 nanoseconds and the 3 dB bandwidth of the power spectrum is approximately 3.2 GHz with maximum power density at 2 GHz.

In order to pass information over a UWB communication link, trains of pulses must be transmitted with some characteristic of a pulse or group of pulses varied in order to distinguish between "0" and "1." The time between consecutive pulses should be determined in a pseudo-random manner in order to smooth the energy spikes in the frequency spectrum. Reception of the transmitted pulse train is done by correlating the received signal with a similar sequence of pulses generated in the receiver. A large number of communication links can be maintained simultaneously and independently by using different pseudo-random sequences for each link.

A pulse similar to that of Figure 5.2-27 can be generated by applying an impulse, or perhaps more conveniently a step-voltage or current, to a linear band limited

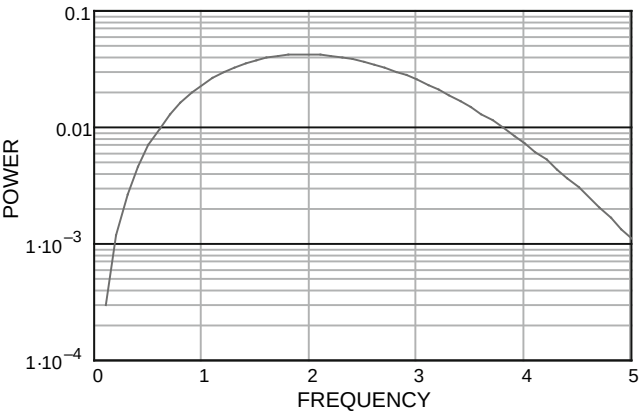


Figure 5.2-28 Spectrum of monopulse.

network. Figure 5.2-29 is a simulation of a sequence of UWB pulses created by stimulating a bandpass filter with a pseudo-randomly spaced sequence of impulses. The figure also shows the power spectrum of that sequence. The network that creates the individual UWB pulses includes the transmitter antenna, the propagation channel, and the receiving antenna, whose characteristics, in terms of impulse response or amplitude and phase vs. frequency must be known and accounted for in designing the system.

There are several ways of representing a UWB pulse as "1" or "0." One method is to advance or retard the transmitted pulse with respect to the expected time of arrival of the pulse in the receiver according to the agreed pseudo-random time sequence. Another method is to send the pulse with or without inversion. In both cases the correlation of the received pulse with a "template" pulse generated in the receiver will result in a different polarity, depending on whether a "1" or a "0" was transmitted.

Detection of UWB bits is illustrated in Figure 5.2-30. A "1" monopulse is represented by a negative line followed by a positive line, and a "0" monopulse by the inverse—a positive line and a negative line. The synchronized sequence generated in the receiver is drawn on the second line and below it the result of the correlation operation $\int f(t) \cdot g(t)dt$ where $f(t)$ is the received signal and $g(t)$ is the locally generated sequence. By sampling this output at the end of each bit period and then resetting the correlator, the transmitted sequence is reconstructed in the receiver.

As mentioned above, an individual bit can be represented by more than one sequential monopulse. Doing so increases the processing gain by the number of monopulses per bit. Processing gain is also an inverse function of the pulse duty cycle. This is because, for constant average power, the power in the pulses contributing to each bit must be raised by $(1/duty\text{-}cycle)$. By gating out the noise except during the interval of the expected incoming pulse, the signal-to-noise ratio will only be a function of the power in the pulse, regardless of the duty cycle. An example may make the explanation clearer. Let's say we are sending data a rate of 10 Mbps. A UWB pulse in the transmitter is 200 picoseconds wide; 20 pulses represent one bit. The time between bits is $1/(10*10^6) = 100$ nanoseconds. So the time between pulses is (100 ns) /20 = 5 ns. The duty cycle is (200 ps)/(5 ns) = 25. Now the processing gain attributed to the number of UWB

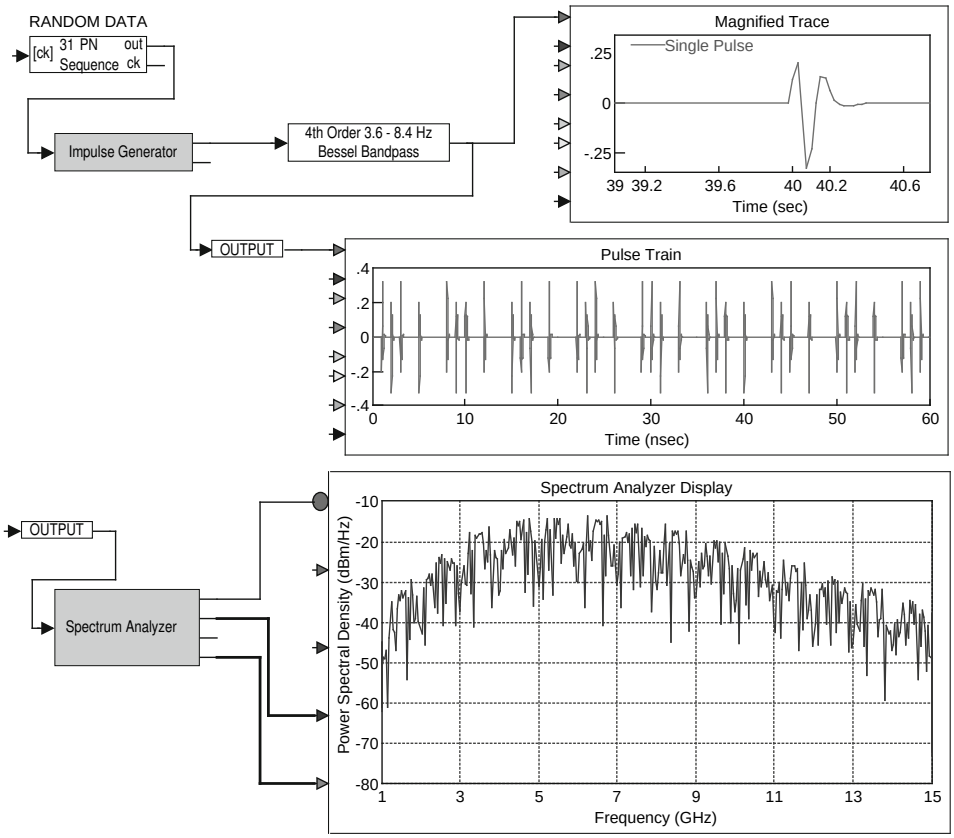


Figure 5.2-29 Simulated sequence of UWB pulses.

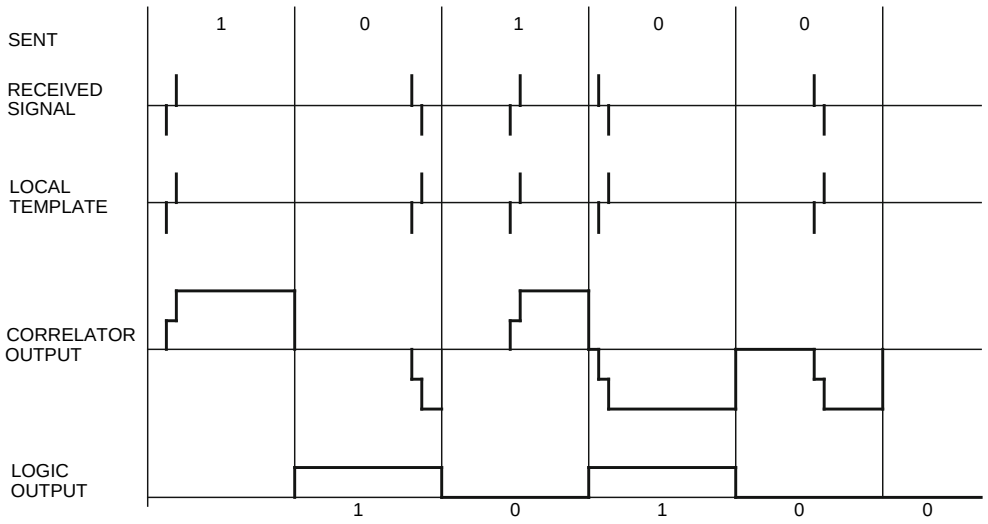


Figure 5.2-30 Detection of UWB bit sequence.

pulses per bit is $10\log(20) = 13$ dB. That due to the duty cycle is $10\log(25) = 14$ dB. Total processing gain is $13 + 14 = 27$ dB.

A simplified block diagram of a UWB system is shown in Figure 5.2-31. A key to the generation of UWB pulses is the ability to create short impulse or step functions with rise times on the order of tens or at the most hundreds of picoseconds, and to detect the UWB pulses that result from their application. High speed integrated circuits can be employed or special circuit elements, such as tunnel diodes or step recovery diodes, can be incorporated.

The main FCC regulations pertaining to ultra-wideband transmissions are described elsewhere. Conditions for using UWB in Europe are also being considered by the European Union. Due to the fear of interference with vital wireless services in the wide bandwidths covered by UWB radiation, spectral density limits allowed by the FCC and presumably to be permitted in the European Union are relatively low, on the order of spurious radiation limits for conventional unlicensed transmissions. However, as we have seen, high-processing gains can be achieved with UWB and communication ranges on the order of tens or hundreds of meters at high-data rates can be expected. Also, the FCC has indicated its intention to monitor the effects of UWB transmissions on other services, once equipment has been put into service in significant quantities, and the agency may be expected to

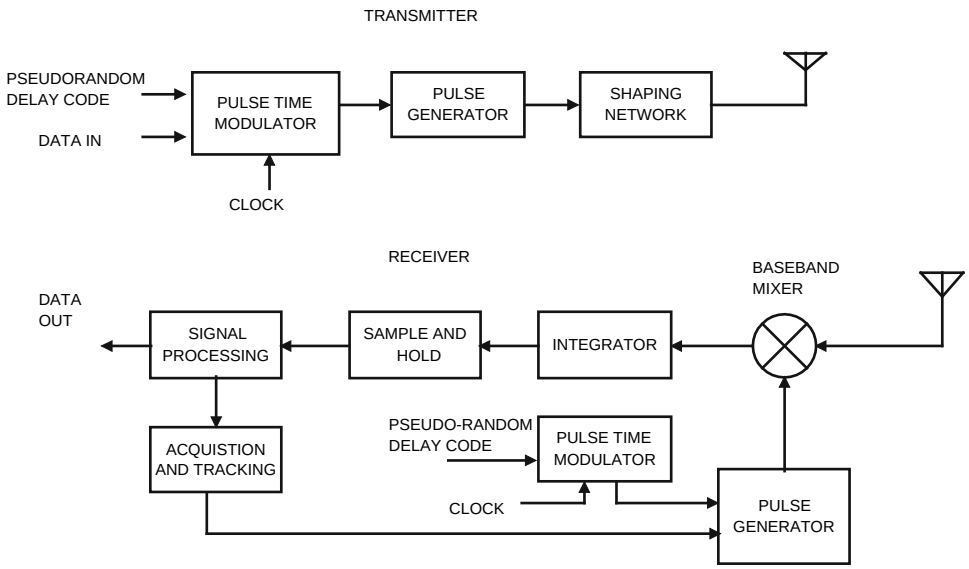


Figure 5.2-31 UWB simplified block diagram.

modify or make its limits more lenient if interference is found not to be a problem. In any case, the unique characteristics of UWB are attractive enough to make this technology an important part of the offerings for short-range wireless communication in the years to come.

5.2.6 Summary

The annually increasing volumes for Bluetooth and Wi-Fi products, stimulated in a large part by the acceptance of industrial standards by the major manufacturers, are causing prices to fall on complex integrated circuits as well as the basic RF components. This trend will open the way for the use of these parts in other short-range applications such as security and medical call systems. Use of sophisticated and proven two-way hardware and link protocols for these and other technically "low end" applications will open them up to much higher usage than they now command. A basic impediment to wireless use will still remain, however, and that is the problem of battery replacement. Reduced voltage and power consumption for integrated circuits will help, as will sophisticated wake-up protocols as are already built-in to Bluetooth. Range limitations may have to be dealt with by a greater use of repeaters than common in today's systems. Another area where advancements are affecting short-range radio is antennas. Both the use of higher frequencies and new designs are reducing antenna size and eliminating a visual reminder of the difference between wired and wireless devices.

The unconventional ultra-wideband technology, since its approval by the FCC, is opening up new civilian applications for short-range wireless, notably in the areas of distance measurement, concealed object location, and high precision positioning systems. Because of its high-interference immunity, and its property of not causing interference, it may successfully compete with and complement other technologies used for short-range radio applications such as personal communications systems, security sensors, and RFID tags.

In summary, advances in short-range radio communication developments in one area feeds its expansion in other areas. Overall, short-range radio will continue to play a major part in the ongoing communication revolution.

This page intentionally left blank

Section Six

Optical data communication

This page intentionally left blank

Chapter 6.1

6.1

Optical fiber, cable and connectors

Ulf L. Österberg

6.1.1 Light propagation

6.1.1.1 Rays and electromagnetic mode theory

Light is most accurately described as a vectorial electromagnetic wave. Fortunately, this complex description of light is often not necessary for a satisfactory treatment of many important engineering applications.

In the case of optical fibers used for tele- and data communication it is sufficient to use a scalar wave approximation to describe light propagation in single-mode fibers and a ray approximation for light propagation in multimode fibers.

For the ray approximation to be valid the diameter of the light beam has to be much larger than the wavelength. In the wave picture we will assume a harmonically time-varying wave propagating in the z direction with phase constant β. The electric field can be expressed as

$$E = E_0(x,y)\cos(\omega t - \beta z). \tag{6.1.1}$$

This is more conveniently expressed in the phasor formalism as

$$E = E_0(x,y)e^{j(\omega t-\beta z)}, \tag{6.1.2}$$

where the real part of the right-hand side is assumed.

A wave's propagation in a medium is governed by the wave equation. For the particular wave in Eq. (6.1.2) the wave equation for the electric z component is

$$\nabla_t^2 E_z(x,y) + \beta_t^2 E_z(x,y) = 0, \tag{6.1.3}$$

where we have introduced

$$\nabla_t^2 = \frac{\partial^2}{\partial_{x^2}} + \frac{\partial^2}{\partial_{y^2}} \quad \text{[Transverse Laplacian]},$$

$$\beta_t^2 = k^2 n^2 - \beta^2 \quad \text{[Transverse phase constant]},$$

$$k = \frac{2\pi}{\lambda} \quad \text{[Free space wave vector]},$$

$$n(x,\ y) \quad \text{[Refractive index]}.$$

The variable kn corresponds to the phase constant for a plane wave propagating in a medium with refractive index n. There is an equivalent wave equation to Eq. (6.1.3) for the H_z component. We have to solve only the wave equation for the longitudinal components E_z and H_z. The reason for this is that E_x and E_y can both be calculated from E_z and H_z using Maxwell's equations.

6.1.1.2 Single-mode fiber

In an infinitely large isotropic and homogeneous medium a light wave can propagate as a plane wave and the phase constant for the plane wave can take on any value, limited only by the available frequencies of the light itself. When light is confined to a specific region in space, boundary conditions imposed on the light will restrict the phase constant β to a limited set of values. Each possible phase constant β represents a mode. In other words, when light is confined it can propagate only in a limited number of ways.

For an engineer it is important to find out how many modes can propagate in the fiber, what their phase

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2002 Elsevier Inc. All rights reserved

constants are, and their spatial transverse profile. To do this we have to solve Eq. (6.1.3) for a typical fiber geometry (Fig. 6.1-1). Because of the inherent cylindrical geometry of an optical fiber, Eq. (6.1.3) is transformed into cylindrical coordinates and the modes of spatial dependence are described with the coordinates r, ϕ, and z. Because the solution is dependent on the specific refractive index profile, it has to be specified. In Fig. 6.1-2 the most common refractive index profiles are shown. For step-index profile in Fig. 6.1-2c, a complete analytical set of solutions can be given [3]. These solutions can be grouped into three different types of modes: TE, TM, and hybrid modes, of which the hybrid modes are further separated into EH and HE modes. It turns out that for typical fibers used in tele- and data communication the refractive index difference between core and cladding, $n_1 - n_2$, is so small ($\sim$0.002–0.008) that most of the TE, TM, and hybrid modes are degenerate and it is sufficient to use a single notation for all these modes—the LP notation. An LP mode is referred to as $LP_{\ell m}$, where the ℓ and m subscripts are related to the number of radial and azimuthal zeros of a particular mode. The fundamental mode, and the only one propagating in a single-mode fiber, is the LP_{01} mode. This mode is shown in Fig. 6.1-3. To quickly figure out if a particular LP mode will propagate, it is very useful to define two dimensionless parameters, V and b.

$$V = ka\sqrt{n_1^2 - n_2^2} \approx \frac{2\pi}{\lambda} \cdot a \cdot n_1 \sqrt{2\Delta}, \tag{6.1.4}$$

where a is the core radius, λ is the wavelength of light, and $\Delta \approx (n_1 - n_2)/n_1$. The V number is sometimes called the normalized frequency.

The normalized propagation constant b is defined as

$$b = \frac{(\beta^2/k^2) - n_2^2}{n_1^2 - n_2^2}, \tag{6.1.5}$$

where b is the phase constant of the particular LP mode, k is the propagation constant in vacuum, and n_1 and n_2 are the core and cladding refractive indexes, respectively.

Equation (6.1.5) is very cumbersome to use because b has to be calculated from Eq. (6.1.3). For LP modes

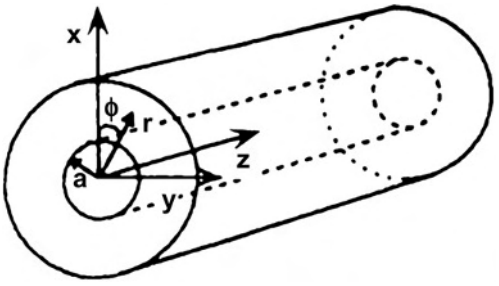


Fig 6.1-1 Typical fiber geometry. Reprinted from Ref. [1], p. 12, courtesy of Academic Press.

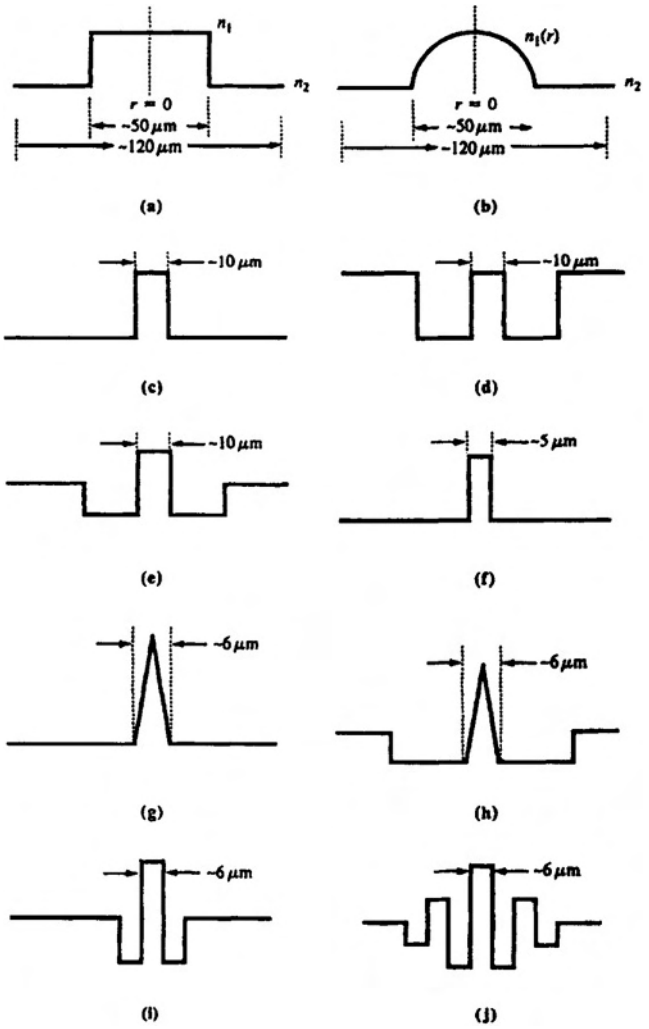


Fig. 6.1-2 Refractive index profiles of (a) step-index multimode fibers, (b) graded-index multimode fibers, (c) match-cladding single-mode fibers, (d, e) depressed-cladding, single-mode fibers, (f–h) dispersion-shifted fibers, and (i, j) dispersion-flattened fibers. Reprinted from Ref. [2], p. 125, courtesy of Irwin.

Marcuse *et al.* [4] have shown that to a very good accuracy the following formulas can be used to calculate b for different

LP_{lm} modes:

$$LP_{01} : b_{01} = 1 - \left[\frac{1+\sqrt{2}}{1+(4+V^4)^{\frac{1}{4}}}\right]^2 \tag{6.1.6}$$

$$LP_{\ell m} \quad (\ell m \neq 01) : b_{\ell m} = 1 - \frac{U_c^2}{V^2} \times \exp\left[\frac{2}{S}\left(\arcsin\left(\frac{S}{U_c}\right) - \arcsin\left(\frac{S}{V}\right)\right)\right]$$

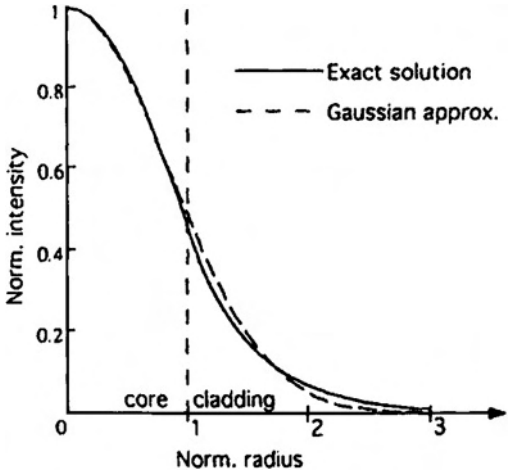


Fig. 6.1-3 Comparison between exact and Gaussian approximation of the radial mode field distribution for the lowest order mode. Reprinted from Ref. [1], p. 14, courtesy of Academic Press.

$$S = \sqrt{U_c^2 - \ell^2 - 1} \tag{6.1.7}$$

$$U_c = A - \frac{B-1}{8A} - \frac{4(B-1)(7B-31)}{3(8A)^3}$$

$$A = \pi \cdot \left[m + \frac{1}{2}(\ell - 1) - \frac{1}{4} \right]; \qquad B = 4(\ell - 1)^2.$$

The graphs in Fig. 6.1-4 were generated using Eqs (6.1.6) and (6.1.7). The normalized propagation constant b can vary only between 0 and 1 for guided modes; this corresponds to

$$n_2 k < \beta < n_1 k. \tag{6.1.8}$$

The wavelength for which b is zero is called the cutoff wavelength, i.e.,

$$b_{\ell m}(V_{co}) = 0 \Rightarrow \lambda_{co} = \frac{2\pi}{V_{co}} \cdot a \cdot n_1 \sqrt{2\Delta}. \tag{6.1.9}$$

Therefore, for wavelengths longer than the cutoff wavelength the mode cannot propagate in the optical fiber.

Cutoff values for the V number for a few LP modes are given in Table 6.1-1. The fundamental mode can, to better than 96% accuracy, be described using a Gaussian function

$$E(r) = E_0 \exp\left[-\left(\frac{r}{w_g} \right)^2 \right], \tag{6.1.10}$$

where E_0 is the amplitude and $2w_g$ is the mode field diameter (MFD) (Fig. 6.1-3). The meaning of the MFD is shown in Fig. 6.1-5. The MFD for the fundamental

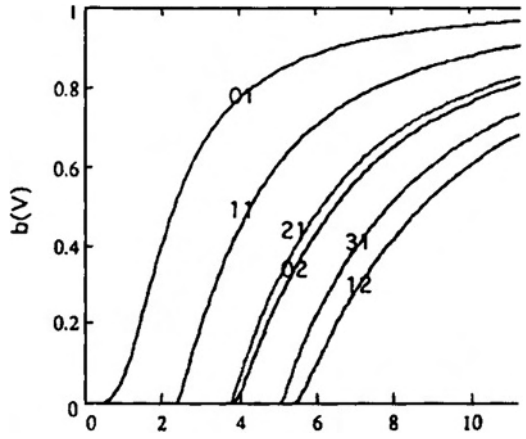


Fig. 6.1-4 Cutoff frequencies for the lowest order LP modes. Reprinted from Ref. [1], p. 15, courtesy of Academic Press.

mode is larger than the geometrical diameter in a single-mode (SM) fiber and much smaller than the geometrical diameter in a multimode (MM) fiber. The optimum MFD is given by the following formula [7]:

$$\frac{w_g}{a} = 0.65 + 1.619 \cdot V^{-3/2} + 2.87 \cdot V^{-6}, \tag{6.1.11}$$

where a is the core radius. Equation (6.1.11) is valid for wavelengths between 0.8 λ_{co} and 2 λ_{co}.

If the radial distribution for higher order modes is needed, it is necessary to use the Bessel functions [3]. In Fig. 6.1-6 the radial intensity distribution is shown for five LP modes in a fiber with $V = 8$. Recommended specifications for a single-mode fiber are summarized in Table 6.1-2.

Table 6.1-1 Cutoff frequencies of various $LP_{\ell m}$ modes in a step index fiber[a].

Mode	V_c	Mode	V_c
$\ell = 0$ modes	$J_1(V_c) = 0$	$\ell = 1$ modes	$J_0(V_c) = 0$
LP_{01}	0	LP_{11}	2.4048
LP_{02}	3.8317	LP_{12}	5.5201
LP_{03}	7.0156	LP_{13}	8.6537
LP_{04}	10.1735	LP_{14}	11.7915
$\ell = 2$ modes	$J_1(V_c) = 0$; $V_c \neq 0$	$\ell = 3$ modes	$J_2(V_c) = 0$; $V_c \neq 0$
LP_{21}	3.8317	LP_{31}	5.1356
LP_{22}	7.0156	LP_{32}	8.4172
LP_{23}	10.1735	LP_{33}	11.6198
LP_{24}	13.3237	LP_{34}	14.7960

[a] Reprinted from Ref. [5], p. 380, courtesy of Cambridge University Press.

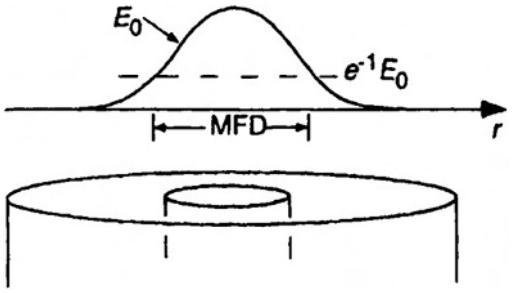


Fig. 6.1-5 The electric field of the HE_{11} mode is transverse and approximately Gaussian. The mode field diameter is determined by the points where the power is down by e^{-2} or where the amplitude is down by e^{-1}. The MFD is not necessarily the same dimension as the core. Reprinted from Ref. [6], p. 144, courtesy of Irwin.

6.1.1.3 Multimode fiber

The previous discussion has in principle been for a step-index MM fiber. Because of the severe differences in propagation time between different modes in a step-index fiber, these are not commonly used in practice.

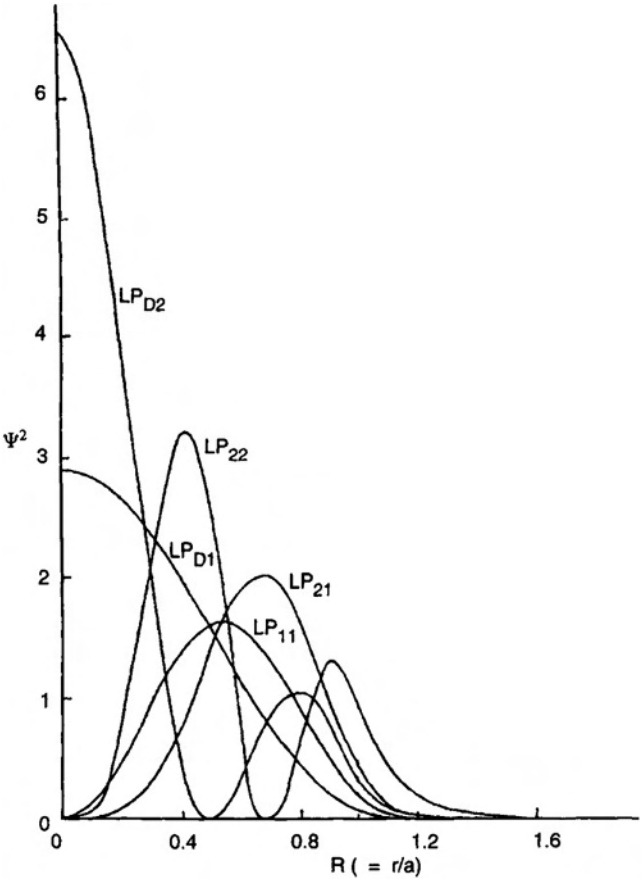


Fig. 6.1-6 Radial intensity distributions (normalized to the same power) of some low-order modes in a step index fiber for $V = 8$. Notice that the higher order modes have a greater fraction of power in the cladding. Reprinted from Ref. [5], p. 382, courtesy of Cambridge University Press.

Table 6.1-2 CCITT recommendation G.652[a].

Parameters	Specifications
Cladding diameter	125 μm
Mode field diameter	9–10 μm
Cutoff wavelength λ_{co}	1100–1280 nm
1550 nm bend loss	≤1 dB for 100 turns of 7.5 cm diameter
Dispersion	≤3.5 ps/nm · km between 1285 and 1330 nm ≤6 ps/nm · km between 1270 and 1340 nm ≤20 ps/nm · km at 1550 nm
Dispersion slope	≤0.095 ps/nm^2 · km

[a] Reprinted from Ref. [2], p. 126, courtesy of Irwin.

Instead, a graded refractive index core is used for a MM fiber (Fig. 6.1-7).

The various graded-index profiles are generated by

$$n^2(r) = \begin{cases} n_1^2\left(1 - 2\Delta\left(\dfrac{r}{a}\right)^q\right); & 0 \le r \le a \\ n_1^2(1-2\Delta) = n_2^2; & r \ge a, \end{cases} \tag{6.1.12}$$

for different q's, q is called the profile exponent. The optimum profile is the one that gives the minimum dispersion; this occurs for q at slightly less than 2.

The total number of modes that can propagate in a MM fiber is given by

$$N = \frac{1}{2}\frac{q}{q+2}V^2. \tag{6.1.13}$$

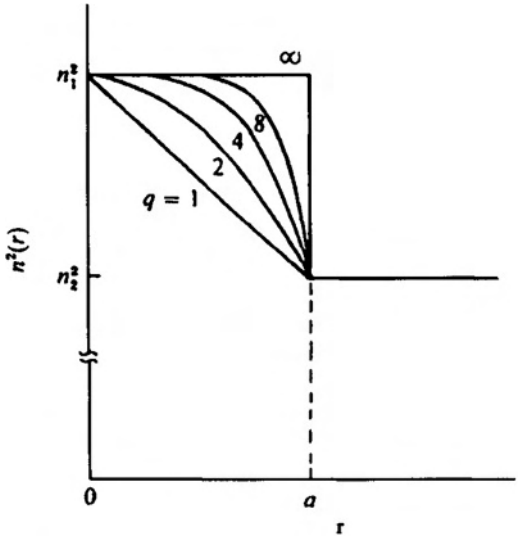


Fig. 6.1-7 The refractive index variation for a power law profile for different values of q. Reprinted from Ref. [5], p. 396, courtesy of Cambridge University Press.

For a step-index fiber $q = \infty$ and $N = V^2/2$. Equation (6.1.13) is only valid for large V numbers.

One can calculate the different ray paths that are possible in a MM fiber using a geometrical optics approach [8]. A more accurate view of the different modes can be obtained by solving Eq. (6.1.3) using the so-called WKB approximation [3]. Using this analysis the phase constants for the different modes can be shown to obey the following relationship:

$$\beta_m = nk\sqrt{1 - 2\Delta\left(\frac{m}{n}\right)^{q/q+2}}, \quad m = 1, 2, \ldots, N. \tag{6.1.14}$$

6.1.1.4 Optical coupling

A first approach to estimate how much light can be coupled into an optical fiber is to use the ray picture. In this picture the light is confined within the core if it undergoes total internal reflection at the core–cladding boundary. This will occur only for light entering the fiber within an acceptance cone defined by the angle θ (Fig. 6.1-8). Rather than stating the angle θ for an optical fiber, it is the convention to give $\sin\theta$, which is called the numerical aperture (NA). The NA is defined as

$$\text{NA} = n \cdot \sin\theta, \tag{6.1.15}$$

where n is the refractive index of the medium the light is coming from. In the case of coupling into an optical fiber, light is usually coming from air and subsequently $n \approx 1$. A more useful formula for the NA can be obtained if we use the dimensionless parameter Δ,

$$\text{NA} \approx n_1\sqrt{2\Delta}. \tag{6.1.16}$$

For an incoherent light source such as a light emitting diode (LED) one can show that the total power accepted by the fiber is given by [9]

$$P \approx \pi \cdot B \cdot A_{\text{fiber}} \cdot (\text{NA})^2, \tag{6.1.17}$$

where B is the LED's radiance (units for radiance is watts per area and steradian).

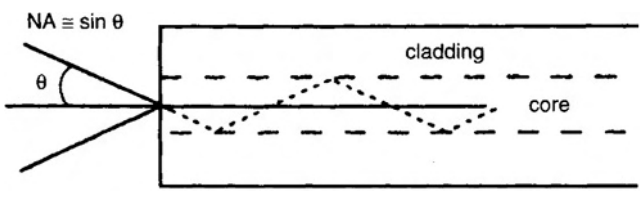


Fig. 6.1-8 Acceptance angle for an optical fiber. Reprinted from Ref. [1], p. 10, courtesy of Academic Press.

It is more common to give a coupling efficiency; thus, giving the total power accepted by the fiber, the efficiency is defined as [10]

$$\eta = \frac{P_{\ell m}}{P_{in}}, \tag{6.1.18}$$

where P_{in} is the power launched into the fiber and $P_{\ell m}$ the power accepted by the $\text{LP}_{\ell m}$ mode. For link budget analyses it is more convenient to deal with coupling losses in units of decibels α:

$$\alpha = 10 \cdot \log\eta. \tag{6.1.19}$$

A more general definition of the coupling efficiency η for an optical fiber that obeys the weekly guiding approximation can be given as [10]

$$\eta = \left| \frac{n_2}{Z_0} \int_A \int_\infty E_i \cdot E^*_{\ell m} dA \right|^2, \tag{6.1.20}$$

where n_2 is the refractive index of the cladding, Z_0 is the free space wave impedance, and E_i and $E_{\ell m}$ are the electric field amplitudes for the incoming light and for the light propagating in mode $\text{LP}_{\ell m}$ in the fiber, respectively.

Coherent light from a laser can often be approximated with a Gaussian beam; furthermore, if we restrict ourselves to a SM fiber, so that the LP_{01} mode can also be approximated as a Gaussian field, it is possible to calculate η analytically [7, 11]:

$$\eta = \left(\frac{4D}{B}\right) e^{-AC/B}, \tag{6.1.21}$$

where

$$A = \frac{(k's_1)^2}{2}$$

$$B = G^2 + (D+1)^2$$

$$C = (D+1)F^2 + 2DFG\sin\theta + D(G^2 + D + 1))\sin^2\theta$$

$$D = \left(\frac{s_2}{s_1}\right)^2$$

$$F = \frac{2\Delta}{k's_1^2}$$

$$G = \frac{2\Delta Z}{k's_1^2}$$

$$k' = \frac{2\pi n_o}{\lambda}$$

where ΔZ is the separation distance between source and fiber, Δ is the lateral displacement, θ is the angular displacement, and s_1 and s_2 are the mode field radii or spot

sizes of the source field and fiber field, respectively. The refractive index of the medium between source and fiber is denoted n_o.

Equation (6.1.21) takes into account four different coupling cases at once (Fig. 6.1-9). If only one of these different coupling cases is present at a time Eq. (6.1.21) can be simplified to:

Case 1.1. Spot-size mismatch $s_1 \neq s_2$:

$$\eta = \left(\frac{2s_1 s_2}{s_1^2 + s_2^2}\right)^2. \tag{6.1.22}$$

Case 1.2. Transverse offset Δ:

$$\eta = \exp\left(-\left(\frac{\Delta}{s_2}\right)^2\right). \tag{6.1.23}$$

Case 1.3. Longitudinal offset ΔZ:

$$\eta = \frac{1}{1 + (\Delta Z / 2Z_0)^2}, \tag{6.1.24}$$

where Z_0 is the Rayleigh range.

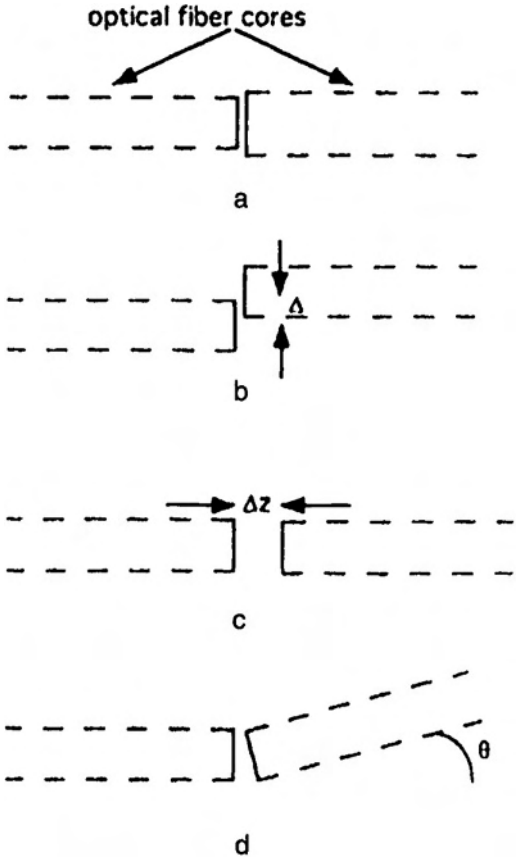


Fig. 6.1-9 Different coupling cases. Reprinted from Ref. [1], p. 18, courtesy of Academic Press.

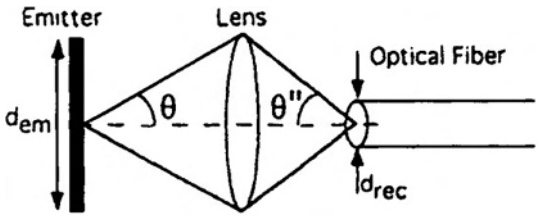


Fig. 6.1-10 Lens for improving coupling efficiency. Reprinted from Ref. [1], p. 20, courtesy of Academic Press.

Case 1.4. Angular misalignment θ:

$$\eta = \exp\left(-\left(\frac{\theta}{\theta_0}\right)^2\right). \tag{6.1.25}$$

If a lens is used in between the emitter and fiber, some modifications to the previous formulas have to be done. What the lens can do for us is to match the output angle of the emitter to the acceptance angle of the receiving fiber. If properly done, the power coupled into the fiber is multiplied with the lens magnification factor M: $M = d_{rec}/d_{em}$, (see Fig. 6.1-10). All the preceding formulas need to be corrected for reflection losses. If the refractive index of the medium between the source and the fiber is denoted n_o, the coupled power into the fiber is reduced with a factor R,

$$R = \left(\frac{n_1 - n_o}{n_1 + n_o}\right)^2. \tag{6.1.26}$$

6.1.2 Optical fiber characterization

6.1.2.1 Optical fiber materials

The material is primarily chosen to provide the minimum attenuation. Table 6.1-3 shows order of magnitude

Table 6.1-3 Scattering loss for several representative glass materials[a].

	dB/km		
Material	**633 nm**	**800 nm**	**1060 nm**
Fused silica	4.8	1.9	0.6
Soda lime	8.5	3.3	1.1
Borosilicate crown	7.7	3.0	1.0
Lead silicate	47.5	18.6	6.0

[a] Reprinted from Ref. [1], p. 21, courtesy of Academic Press.

attenuation at three different wavelengths for four common glass types.

For tele- and data communication fibers, fused silica glass is the preferred material. To provide guiding of the light, the core of the fiber is doped with a few molar percentage of a substance that increases the refractive index. It is also possible to dope the cladding such that its refractive index becomes lower than the pure silica glass index in the core (Fig. 6.1-11).

6.1.2.2 Attenuation

Attenuation is a very important factor in designing effective long-distance fiber optic networks. Consequently, the fabrication methods have improved dramatically during the past 30 years so that attenuation is measured in a few tenths of a dB/km. The dB is defined in Eq. (6.1.19). The various factors affecting the attenuation, in the 0.8- to 1.6-μm wavelength region, are listed in Table 6.1-4. Figure 6.1.12 shows schematically how some of the factors contributing to the overall light attenuation vary with the wavelength in the near-infrared wavelength region. Typical total losses for an optical fiber, at the three different transmission windows at 800, 1300, and 1550 nm, are ≈2 or 3, ≤0.5, and ≤0.2 dB/km, respectively. These numbers are for SM fibers; from Fig. 6.1-13 it can be seen that MM fibers have slightly higher losses. The losses dealt with to date have been due to either intrinsic properties of the glass or extrinsic properties (such as OH and transition metal contents) that come from the particular fabrication method used. In addition, there are bending losses. If the fiber has been improperly cabled or installed these bending losses can be substantial. Bending losses are divided into micro- and macro-bending losses. Micro-bending losses are due to nanometer-size deviations in the fiber. Macro-bending losses are due to visible bends in the fiber. Figure 6.1-14 shows qualitatively how micro- and macrobends contribute to the overall loss in a SM and MM fiber.

Table 6.1-4 Factors affecting attenuation[a].

Intrinsic loss mechanisms
Tail of infrared absorption by Si–O coupling
Tail of ultraviolet absorption due to electron transitions in defects
Rayleigh scattering due to spatial fluctuations of the refractive index
Absorption by impurities
Absorption by molecular vibration of OH
Absorption by transition metals
Structural imperfections
Geometrical nonuniformity at core–cladding boundary
Imperfection at connection or splicing between fibers

[a] From Ref. [3].

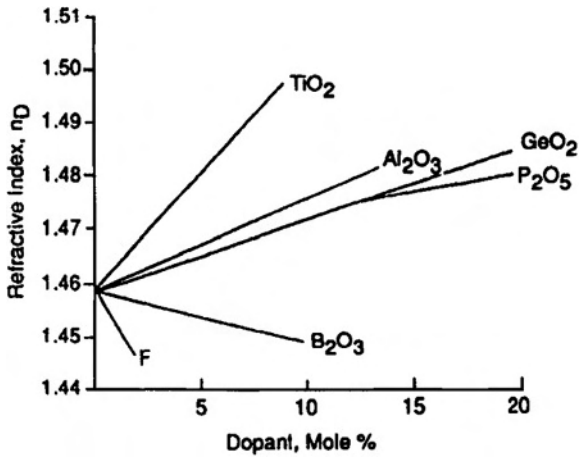


Fig. 6.1-11 Influence of different dopant mole percentages on refractive index of silica glass. Reprinted from Ref. [1], p. 22, courtesy of Academic Press.

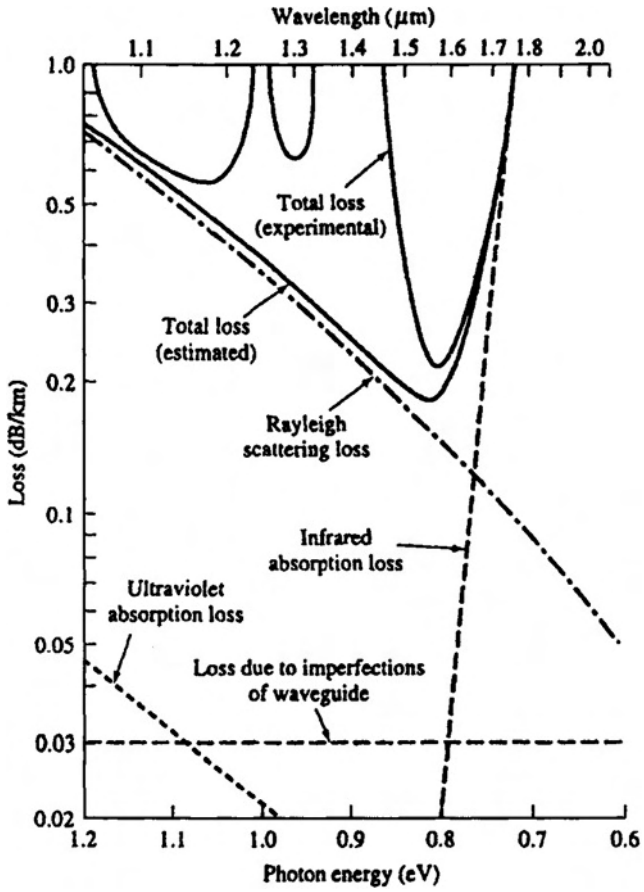


Fig. 6.1-12 Transmission loss in silica-based fibers. Reprinted from Ref. [12], p. 474, courtesy of Irwin.

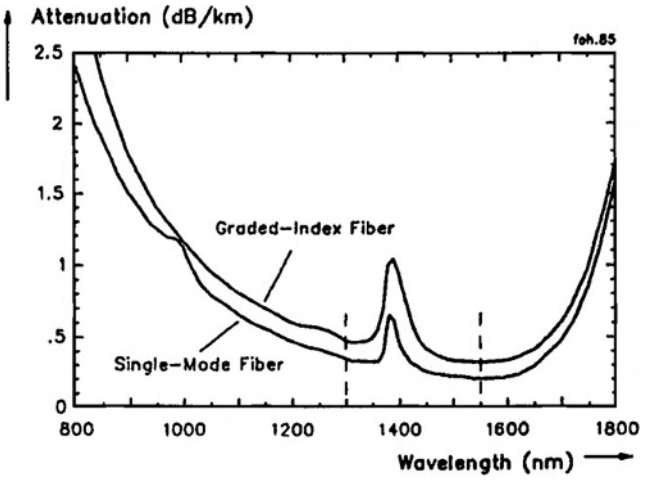


Fig. 6.1-13 Wavelength dependence of fiber attenuation. Reprinted from Ref. [14], p. 8. Copyright © 1989 by Hewlett Packard Company. Reproduced with permission.

6.1.2.3 Dispersion

Dispersion is due to the fact that different wavelengths experience different propagation constants, $\beta(\lambda)$, and therefore travel with different velocities causing a longer temporal pulse at the end of the fiber. Dispersion does not alter the wavelength (frequency) content of the light pulse. From a communications point of view dispersion is a very important factor because it directly affects the bit rate. The following are three major components contributing to the dispersion:

- Material
- Waveguide
- Intermodal

The material and waveguide dispersion are often referred to as chromatic dispersion, measured in units ps/nm · km, which is defined as

$$D_{\text{chr}} = -\frac{1}{L}\frac{dt_g}{d\lambda}, \tag{6.1.27}$$

where L is the fiber length and t_g is the time required to propagate a distance L. The subscript g refers to group velocity. When a pulse propagates in a dispersive medium it propagates with the group velocity,

$$V_g = \frac{d\omega}{d\beta}. \tag{6.1.28}$$

The phase travels with the phase velocity given by

$$V_p = \frac{\omega}{\beta} = \frac{c}{n}. \tag{6.1.29}$$

The dispersion properties are completely determined by the group velocity.

Material dispersion originates from the fact that the refractive index is a function of wavelength, $n(\lambda)$:

$$D_{\text{mat}} = -\frac{\lambda}{c}\frac{d^2 n_1}{d\lambda^2}. \tag{6.1.30}$$

The refractive index can be calculated from the Sellmeier equation:

$$n_1^2(\lambda) = 1 + \sum_{i=1}^{3} \frac{a_i \lambda^2}{\lambda^2 - b_i}, \tag{6.1.31}$$

where the constants a_i and b_i may vary with the particular fiber composition. These constants have been accurately measured and tabulated for different Ge, B and P concentrations [16].

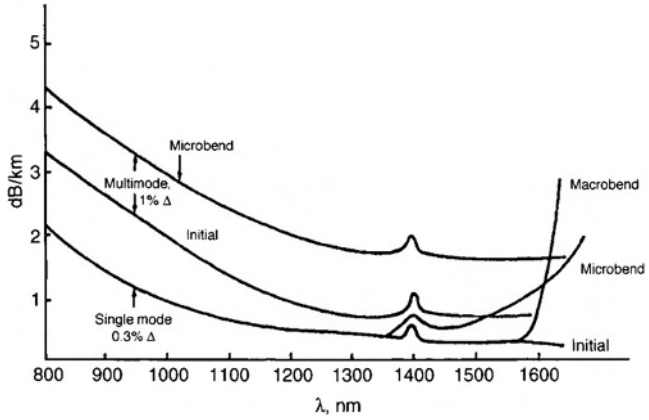


Fig. 6.1-14 Bend-induced losses of optical fibers. Reprinted from Ref. [15], p. 1.33, courtesy of McGraw-Hill.

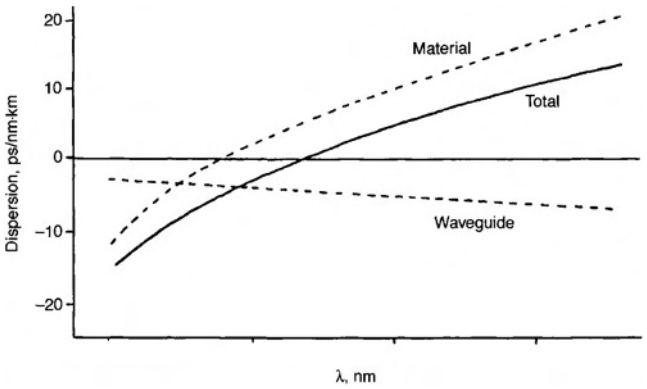


Fig. 6.1-15 Dispersion vs wavelength for single-mode fiber. Reprinted from Ref. [15], p. 1.38, courtesy of McGraw-Hill.

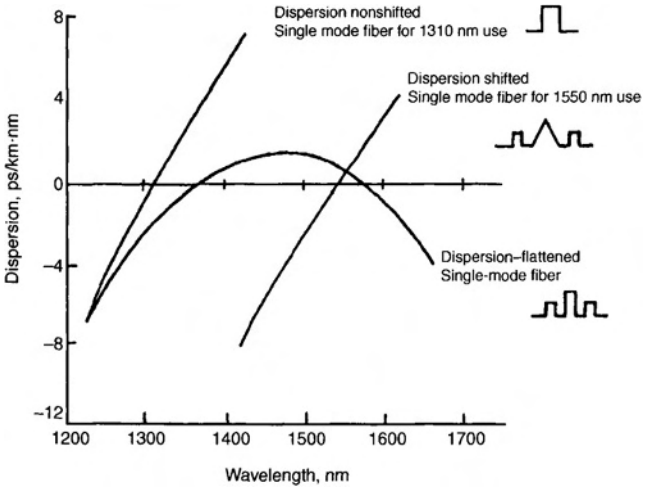


Fig. 6.1-16 Spectral disposition for various single-mode fiber profiles. Reprinted from Ref. [15], p. 1.38, courtesy of McGraw-Hill.

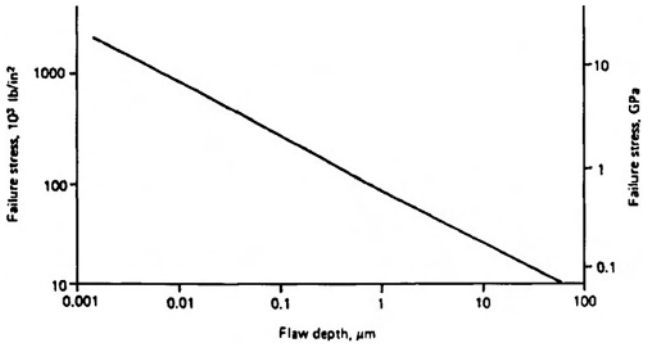


Fig. 6.1-17 Failure stress vs flaw size. Reprinted from Ref. [15], p. 1.40, courtesy of McGraw-Hill.

Waveguide dispersion is due to the different propagation characteristics of the light in the core and cladding. Keep in mind that a large portion of the light (30–40%) travels in the cladding for the LP_{01} mode around the cutoff frequency for the next higher order mode. The waveguide dispersion can be approximated, using the normalized propagation constant b, as

$$D_{\text{wg}} \simeq -\frac{\lambda}{c} \cdot n_1 \cdot \Delta \cdot \frac{d^2 b}{d\lambda^2}. \tag{6.1.32}$$

The total chromatic dispersion as well as its constituents are plotted in Fig. 6.1-15. As shown in Fig. 6.1-2, an optical fiber can have many different refractive index profiles. How the refractive index profile can affect the dispersion properties of the fiber is shown in Fig. 6.1-16.

Intermodal dispersion arises from the different travel times for the different modes in the fiber. For modal dispersion the units are nsec/km and D_{mod} is defined as follows [2]:

$$D_{\text{mod}} = t_{\text{g}}^{\max} - t_{\text{g}}^{\min}, \tag{6.1.33}$$

where $t_{\text{g}}^{\max}$ and $t_{\text{g}}^{\min}$ are the propagation times for the modes traveling the longest and shortest distances, respectively. Using the group index N_1 [3],

$$N_1 = n_1 + k\frac{dn_1}{dk}, \tag{6.1.34}$$

where the term $k\ (dn_1/dk)$ is the material dispersion for a plane wave traveling in the core region; we can approximately write the modal dispersion in a step-index fiber as follows [2]:

$$D_{\text{mod}} \simeq -\frac{N_1}{c}\Delta, \tag{6.1.35}$$

Table 6.1-5 Physical properties of materials[a].

Property	Silica glass	Copper	Aluminium	Steel
Chemical sign	Sio_2	Cu	Al	Fe
Specific gravity (g/cm^3)	2.20	8.9	2.70	7.9
Tensile strength (kg/mm^2)	500	25	10	120
Young's modulus (kg/mm^2)	7200	12,000	6300	20,000
Elongation (%)	2–8	20–30	7–20	5–15
Coefficient of thermal expansion	5×10^{-7}	1.7×10^{-6}	2.3×10^{-5}	1×10^{-5}
Specific heat (cal/°C·g)	0.20	0.09	0.5	0.1

[a] Reprinted from Ref. [18], p. 53, courtesy of Dekker.

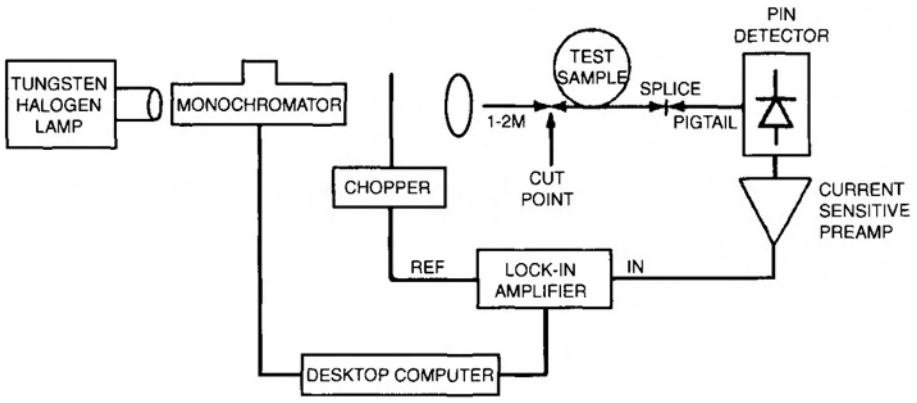


Fig. 6.1-18 Experimental setup for attenuation measurements by the cutback technique. This same apparatus can be used to measure cutoff wavelength. Reprinted from Ref. [20], p. 336, courtesy of Academic Press.

and for a graded-index fiber, with $q = 2(1 - \Delta)$, as

$$D_{\text{mod}} \simeq -\frac{N_1}{c}\frac{\Delta^2}{8}. \tag{6.1.36}$$

Note that for most fibers $N_1 \simeq n_1$ is an excellent approximation [3].

The total fiber disposition in a MM fiber, units nesc/km, can be shown to be the following [2]:

$$D_{\text{tot}}^2 = D_{\text{chr}}^2 \cdot \Delta\lambda^2 + D_{\text{mod}}^2, \tag{6.1.37}$$

where $\Delta\lambda$ is the spectral bandwidth of the light source.

Knowledge of the total dispersion, D_{tot}, can be transmitted into bandwidth (BW), units MHz · km, as follows [17]:

$$\text{BW} \simeq \frac{350}{D_{\text{tot}}}. \tag{6.1.38}$$

6.1.2.4 Mechanical Properties

Optical fibers have outstanding signal characteristics such as low attenuation and small dispersion. For optical fibers to be practically useful they must also have good mechanical properties to withstand environmental strains. In Table 6.1-5, physical properties for different materials used in communication cables are compared.

Silica glass has a very high threshold for breaking when there are no flaws, either on the surface or inside the glass. The tensile strength has been estimated to be 20 GPa [19]. Because optical glass fibers have a large surface-to-volume ratio they are very prone to failure from surface flaws that propagate into the glass. The actual stress for which the fiber breaks as a function of the size of the surface flaw is shown in Fig. 6.1-17.

When fibers are tested for their mechanical reliability, it is common to use short fibers for destructively testing the fibers' breaking threshold for static and dynamic

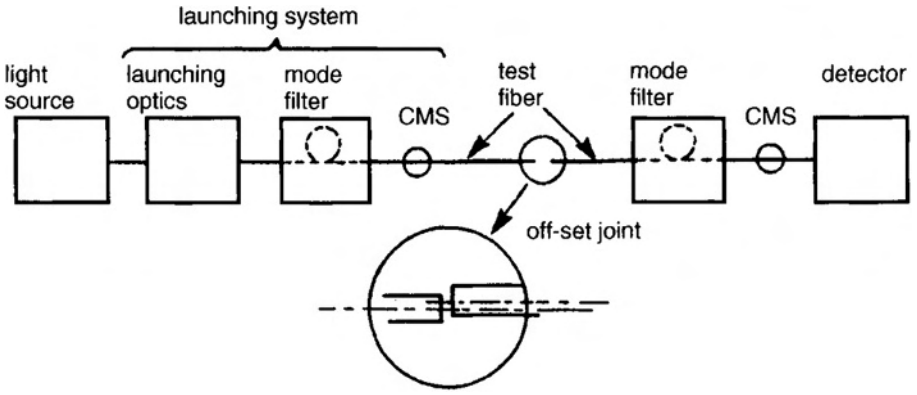


Fig. 6.1-19 Offset technique for mode field measurement—a block diagram; CMS, cladding mode stripper. Reprinted from Ref. [18], p. 426, courtesy of Marcel Dekker Inc.

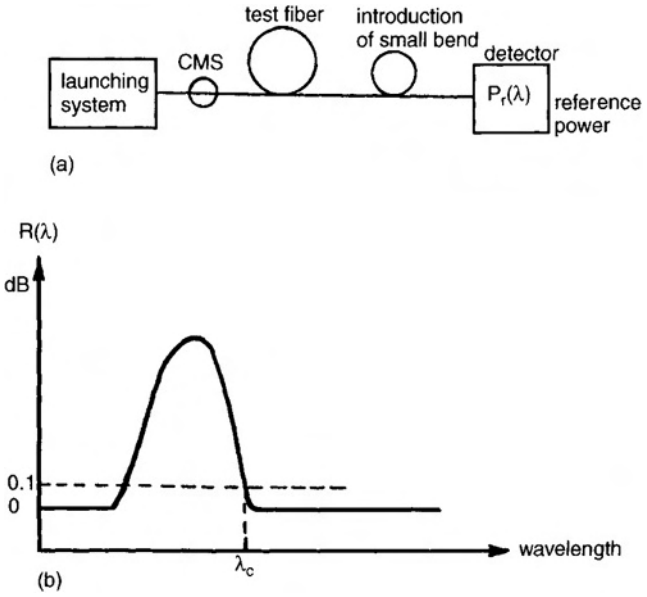


Fig. 6.1-20 Typical setup of cutoff wavelength measurement: (a) bending and (b) typical measurement of graph of the cutoff wavelength, single-mode reference (bending method). Reprinted from Ref. [18], (a) p. 430, and (b) p. 431, courtesy of Marcel Dekker Inc.

loads. Additionally, a screen test is performed on long samples of the fibers to find the effective strength [19].

6.1.2.5 Measurement techniques

In this section I will briefly introduce one measurement method each for measuring attenuation, mode field diameter, cutoff wavelength, and chromatic dispersion.

Attenuation (the cutback method)

The transmission is measured as a function of wavelength for the full-length fiber, $P_L(\lambda)$. The fiber is then cut approximately 1 or 2 m away from the input end and a new transmission measurement is performed, $P_r(\lambda)$ (Fig. 6.1-18). The attenuation constant α is then calculated as follows [20]:

$$\alpha = 10 \cdot \log \frac{P_L(\lambda)}{P_r(\lambda)}. \qquad (6.1.39)$$

Mode Field Diameter (the Near-Field Method)

A small area detector is used to scan the magnified image of the fiber end (Fig. 6.1-19). The experimental data are fitted to Eq. (6.1.10), from which the mode-field diameter can be calculated.

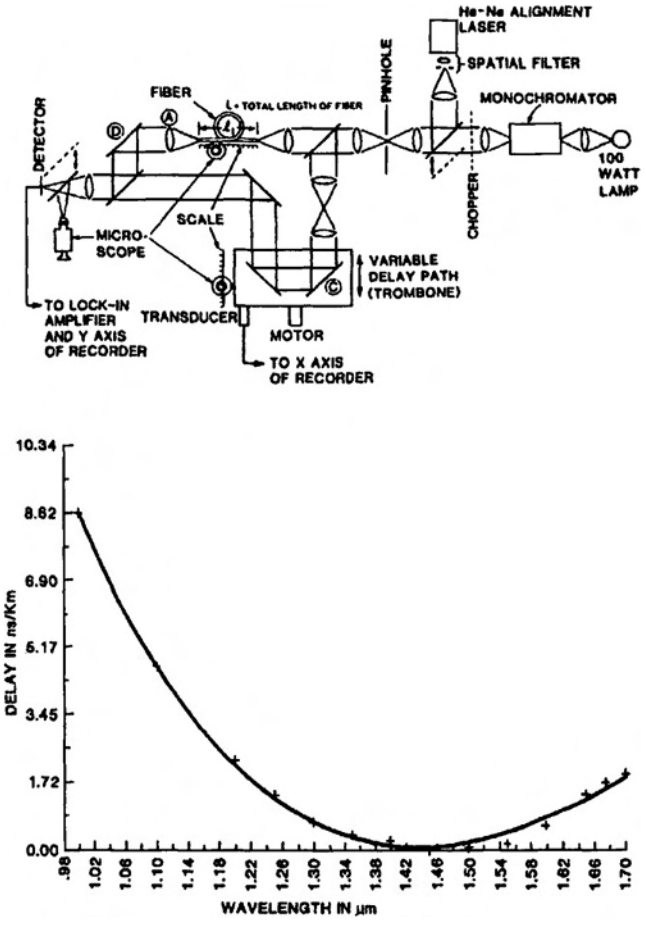


Fig. 6.1-21 Optical arrangement of dispersion measured by the interferrometric technique and the delay curve. Reprinted from Ref. [20], p. 351, courtesy of Academic Press.

Cutoff Wavelength (the Single-Bend Attenuation Method)

In this method the power transmission is measured as a function of wavelength for two different fiber bending situations. In the first measurement a prescribed bend diameter of 28 cm is used (CCITT and EIA standards).

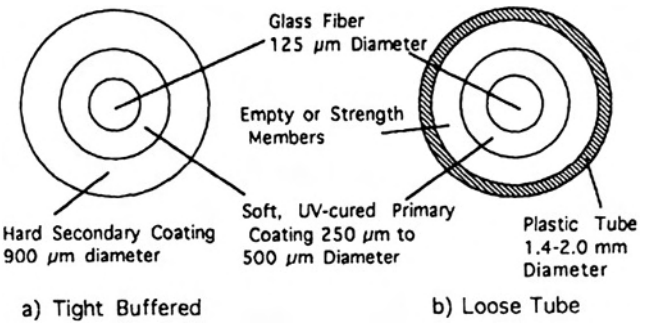


Fig. 6.1-22 Common single-fiber cable types. Reprinted from Ref. [1], p. 36, courtesy of Academic Press.

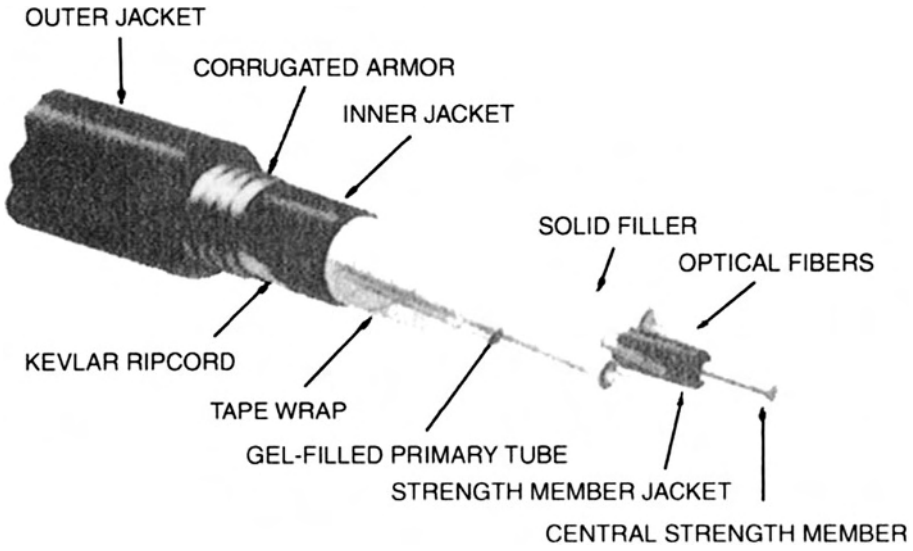


Fig. 6.1-23 Fiber optic cable-corrugated armor. Reprinted from Ref. [21], p. 247, courtesy of Academic Press.

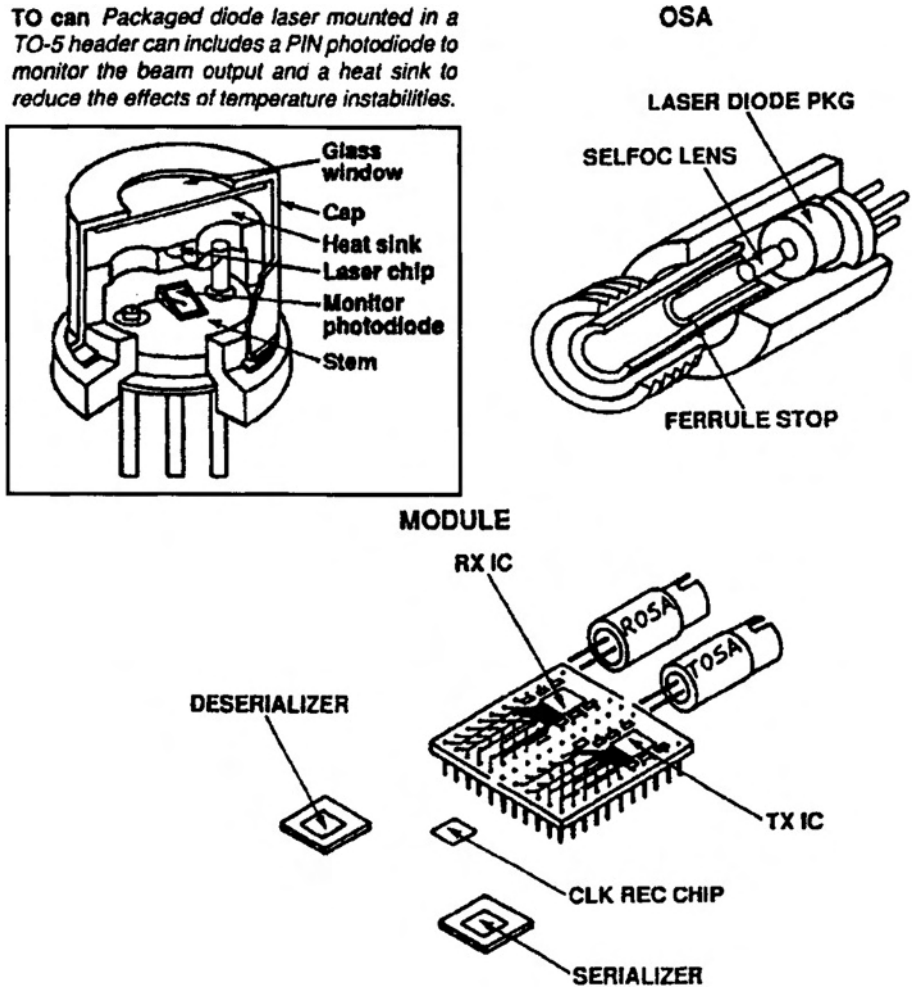


Fig. 6.1-24 Common solution today; Laser TO can → optical subassembly (OSA) → module. There are three levels of packaging and several active alignments. Reprinted from Ref. [22], p. 134, courtesy of Academic Press.

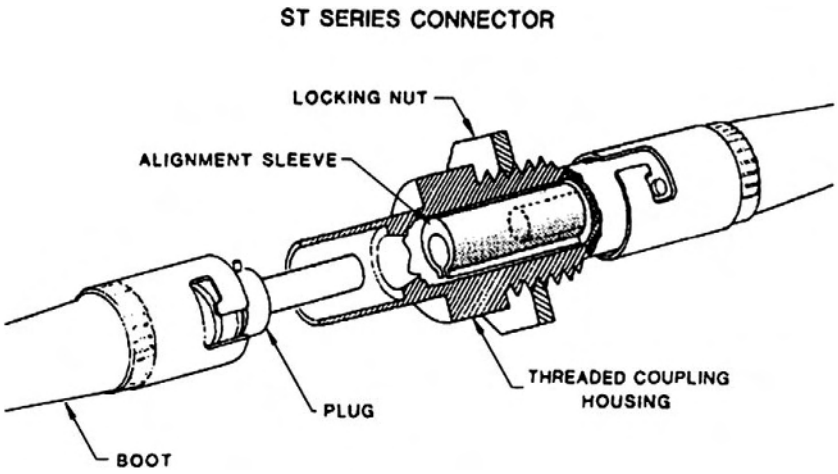


Fig. 6.1-25 Straight tip ferrule single fiber connector. Reprinted from Ref. [23], p. 303, courtesy of Academic Press.

In the second measurement an additional loop with diameter 4–6 cm is used to, as much as possible, attenuate the higher order modes. The cutoff is then defined as the wavelength for which the normalized power becomes linear to within 0.1 dB (Fig. 6.1-20).

Chromatic dispersion (interferometric method)

This method measures the phase delay for different wavelengths. The source is an incoherent high-pressure arc lamp and a typical delay curve is shown in Fig. 6.1-21.

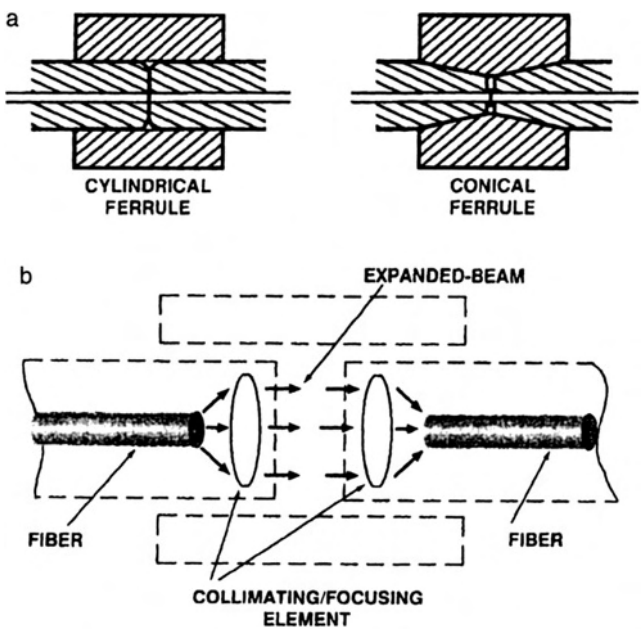


Fig. 6.1-26 (a) Example of two popular alignment mechanisms used in fiber optic connectors, (b) Schematic representation of a typical expanded-beam type connector. Reprinted from Ref. [23], p. 316 (a) and p. 318 (b), courtesy of Academic Press.

6.1.3 Cable designs

Because an optical fiber is very brittle and susceptible to chemical degradation it is important to protect the optical performance. Two common types of fiber optic cable construction are shown in Fig. 6.1-22. Common for all cable types is that the individual fibers are coated with some kind of organic material, examples of which include the following [16]:

- Polydimethyl siloxane
- Silicone oils
- Extrudates
- Acrylates

For ease of manufacturing and to prevent the glass surface from being scratched the coating is applied during the drawing process. Often, the first coating is not

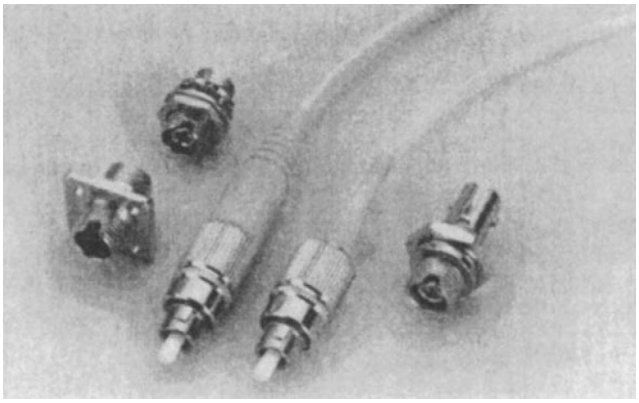


Fig. 6.1-27 FC connectors and adaptors. Reprinted from Ref. [24], p. 16, courtesy of Molex Fiber Optics, Inc.

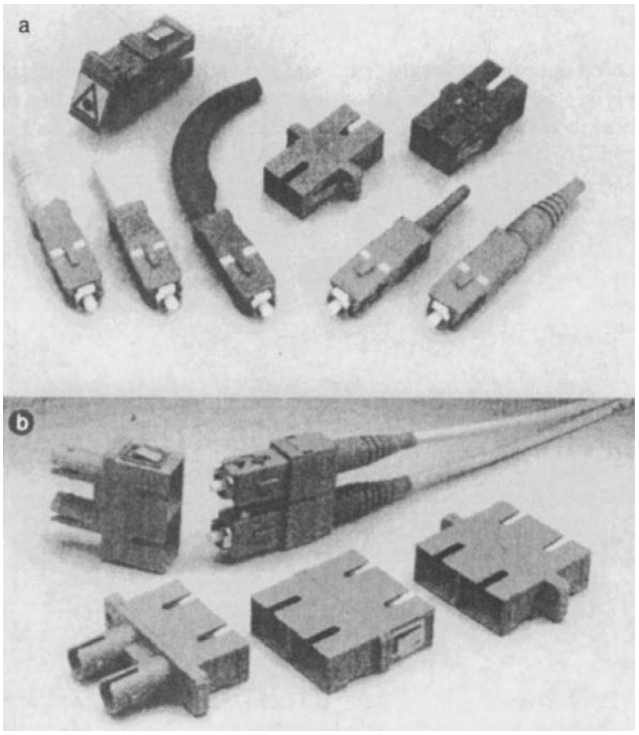


Fig. 6.1-28 (a, b) Spring latches. Reprinted from Ref. [24], p. 24 (a), and p. 34 (b), courtesy of Molex Fiber Optics, Inc.

sufficient to protect the fiber and additional coatings are used, which are commonly referred to as buffering. The buffered fiber is then embedded in various strength members and additional protective jackets. These additional layers have as their primary goals to protect the fibers from water, rodents, and temperature fluctuations and to reduce stress in the fiber (Fig. 6.1-23).

6.1.4 Connectors

6.1.4.1 Connector types

It is important to be able to easily reconfigure communication links for cost and performance optimization, to replace link subunits, and for relocating equipment [1]. To do this, connectors are a key component. There are two types of fiber optic connections:

- Fiber to transceiver
- Fiber to fiber

An example of a laser transceiver module is shown in Fig. 6.1-24. The fiber-to-fiber connectors are divided into simplex (one fiber to one fiber) and duplex (two fibers to two fibers) connectors (Fig. 6.1-25). The two basic connector designs are butt joints and expanded beam joints (Figs. 6.1-26a and 6.1-26b).

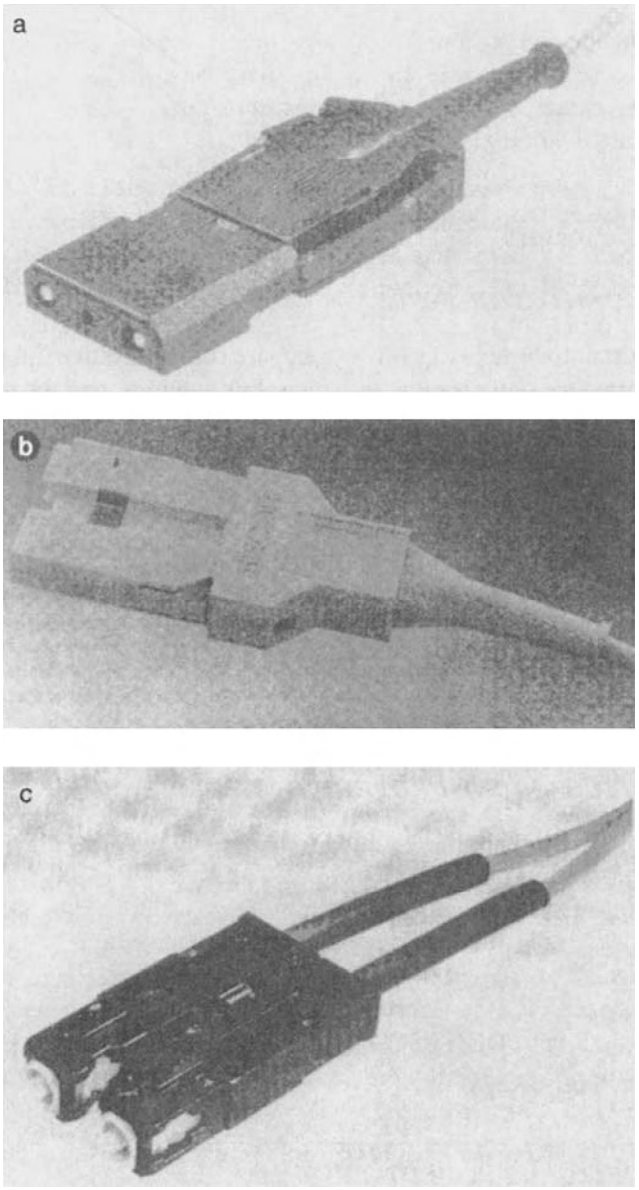


Fig. 6.1-29 Duplex connectors: (a) ESCON (R), (b) FDDI, (c) FCS duplex (fiber channel standard compliant). Reprinted from Ref. [1], p. 50, courtesy of Academic Press.

6.1.4.2 Mechanical properties

Connectors are further differentiated depending on the particular mechanical method used to join the connectors.

FC connectors use threaded fasteners (Fig. 6.1-27). SC connectors use a spring latch (Fig. 6.1-28). These connectors can come as both simplex and duplex.

Variations on the SC connector, developed for specific standardized local area networks, are FDDI, ESCON, and FCS connectors (Figs. 6.1-29a–c). A common

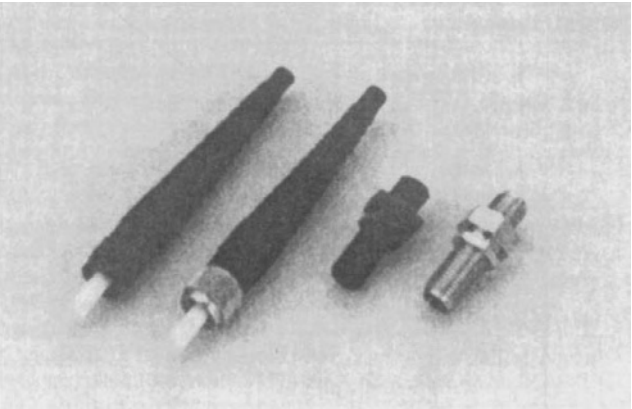


Fig. 6.1-30 SMA connectors and adapters. Reprinted from Ref. [24], p. 42, courtesy of Molex Fiber Optics, Inc.

connector for MM data communication networks is the SMA connector (Fig. 6.1-30).

6.1.4.3 Optical properties

There is obviously some additional loss introduced into a communications link with the incorporation of a connector (Fig. 6.1-31). As can be seen from Fig. 6.1-31 there is both a local and a distributed loss resulting from the connector joint. Contributing connector loss factors are shown in Table 6.1-6. The local loss factors can be calculated using the formulas in Section 6.1.1.4. The distributed losses are due to changes in the model launching conditions [15].

Connector losses can be as high as 0.5 dB (this corresponds to an additional km of fiber at 1.3 μm). With improved manufacturing technology, losses between 0.1 and 0.2 dB are obtained for new connectors. Loss data for a FDDI connector are shown in Fig. 6.1-32.

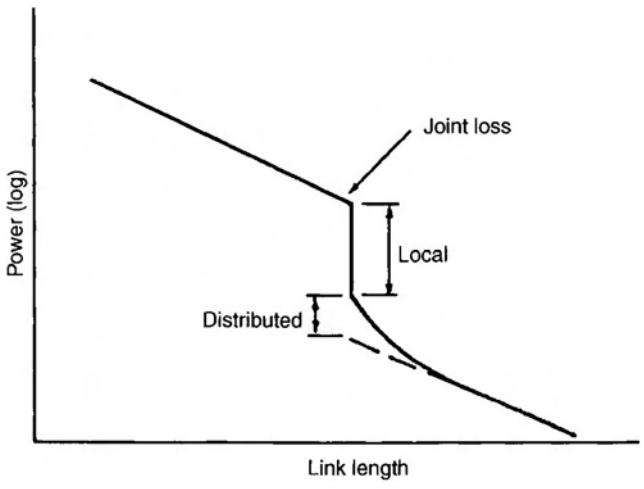


Fig. 6.1-31 Local and distributed (transient) losses and their launch dependence. Reprinted from Ref. [15], p. 4.40, courtesy of McGraw-Hill.

Table 6.1-6 Connector loss factors[a].

Extrinsic	Intrinsic
Lateral offset	Mode field diameter mismatch
Axial gap	Index profile mismatch
Tilt	Core eccentricity
End-face quality	Fiber and ferrule hole clearance
Reflections	Ferrule hole eccentricity

[a] Reprinted from Ref. [1], p. 55, courtesy of Academic Press.

6.1.5 Optical fiber bragg gratings

An optical glass fiber has many advantages compared to other materials such as a copper cable. One of these advantages is its inertness to external changes such as heat, pressure, electric and magnetic fields, toxicity, etc. However, the glass in an optical fiber can be altered by intense laser light [25, 26]. This fact has led to the manufacturing of so-called fiber Bragg gratings (FBG). These in-core fiber gratings have revolutionized optical fiber communications systems, [27, 28]. The FBGs can be used for dispersion compensation [29], filter applications [30], mode conversion [31], gain equalization [32], and wavelength multiplexing [33]. FBGs were first discovered in 1978 by Hill *et al.* [25]. These early gratings were created by the interference between forward and backward traveling waves in the fiber. The backward traveling wave comes from Fresnel reflection at the output end. At the crests of this interference pattern, where the light is the strongest, interactions between the light and Ge defects in the glass alter the refractive index $\Delta n \approx 10^{-3}$–10^{-5}. The periodicity of these gratings is given by

$$\Lambda = \frac{\lambda}{2n_{\text{eff}}} \tag{6.1.40}$$

where n_{eff} is the effective index of the fundamental mode. Evident from equation 6.1.40 is the fact that for a given laser only one periodicity can be produced. Not very practical! Because the interaction between the light and the defects is optimum for a small wavelength region in the UV one cannot just change laser to obtain a new grating periodicity. It was therefore a major

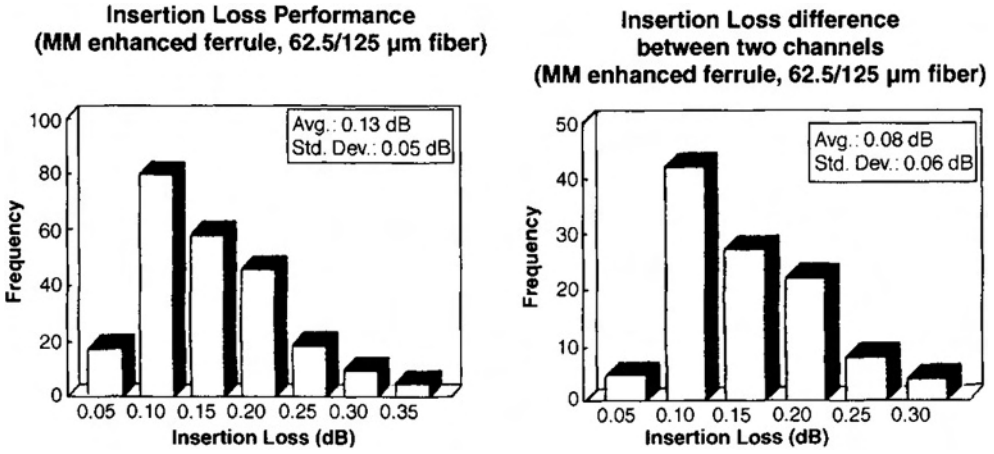


Fig. 6.1-32 Insertion loss performance for FDDI connectors. Reprinted from Ref. [24], p. 46, courtesy of Molex Fiber Optics, Inc.

breakthrough when Meltz *et al.* [34] invented a new holographic method with which gratings of an (almost) arbitrary periodicity can be written into the fiber. This opened up the possibility for making gratings at wavelengths important for tele- and data communications. Further improvements have led to the fabrication of gratings using phase masks [35], a method that allows gratings to be manufactured while the fiber is being drawn.

An optical fiber grating reflects wavelengths at and close to the Bragg wavelength ($\lambda_B = 2n_{eff} \cdot \Lambda$). A typical reflection spectrum is shown in Fig. 6.1-33. The equations used to calculate the spectrum in Fig. 6.1-33 are found in [26]. Reflection takes place because light is partially reflected at each refractive index crest and maximum reflection occurs when each partial reflection is in phase with its neighbors, Fig. 6.1-34.

Let's look at two specific applications of these Bragg gratings. The first one is dispersion compensation. As mentioned in Section 6.1.2.3, the refractive index is a function of wavelength in the glass, so a light pulse will always broaden in time as it propagates through the fiber. This broadening becomes a big problem in high bandwidth time-multiplexed systems because the pulses, if left alone, will eventually merge together and destroy any hope of detecting individual pulses. By using an FBG in which the periodicity varies linearly along the fiber, the grating is said to be chirped, one can

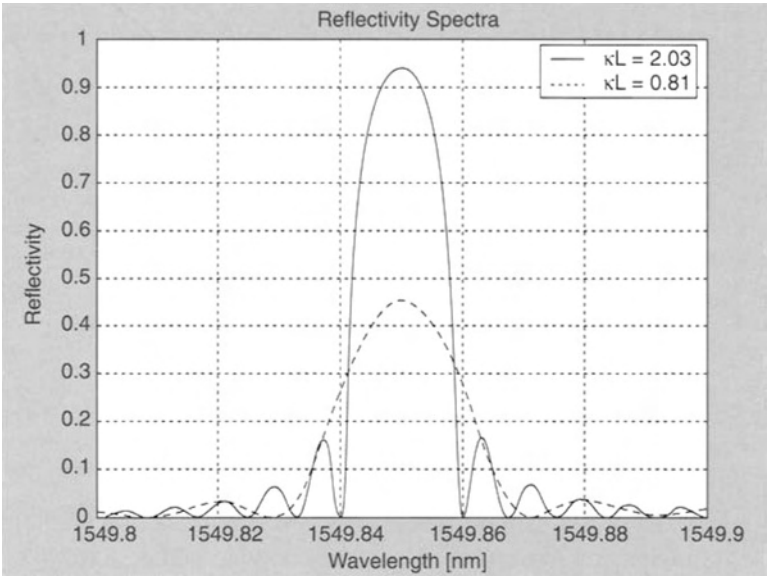


Fig. 6.1-33 Reflectivity spectra for two different values of the induced refractive index change in the core.

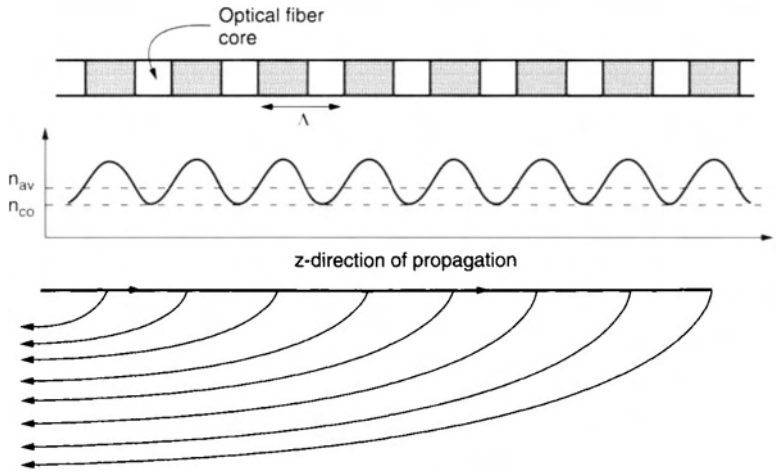


Fig. 6.1-34 Buildup of back-reflected light in an optical fiber grating.

delay (in time) different wavelengths different amounts and thereby sharpen the pulse in time. Dispersion compensating FBGs of up to 1500 ps/nm in the wavelength region around 1550 nm are commercially available.

The second application is concerned with add/drop multiplexer/demultiplexers (mux/demux). One example of such a device is shown in Fig. 6.1-35 [33]. The purpose of such a device is to insert or remove a particular wavelength in a wavelength-division system. For example, to remove one wavelength out of a stream of wavelengths one can use the combination of a Mach-Zehnder interferometer in conjunction with FBGs. Assuming one would want to remove wavelength λ_2, it is necessary to have two identical FBGs, one in each of the two interferometer arms (Fig. 6.1-35) and both optimized for λ_2 (i.e., only λ_2 satisfies the Bragg condition in Eq. 6.1.40). At the input the 4 wavelengths are divided so that 50% of each wavelength goes into each arm. Due to the FBGs λ_2 will be reflected from each arm and the other three wavelengths will be transmitted. The two streams of transmitted light are subsequently combined in the right coupler and interfering in such a way that all the transmitted light will exit through port 4. Similarly, the two portions of λ_2 will be combined and interfere in the left coupler to emerge at port 2.

Due to the inherent symmetry of these balanced Mach-Zehnder interferometers they can simultaneously perform both multiplexing and demultiplexing. A commercial mux/demux device is capable of adding or dropping wavelengths that are < 1 nm apart (< 100 GHz @ 1550 nm) with less than -30 dB crosstalk between different wavelengths.

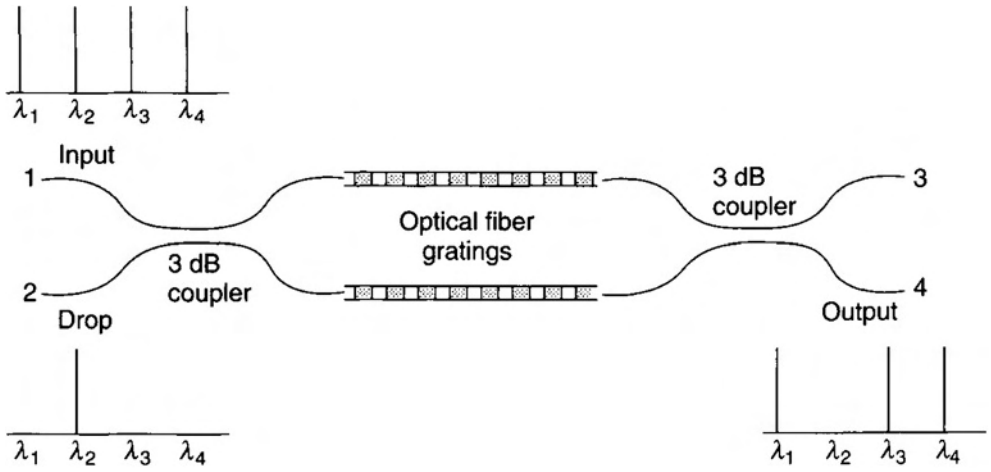


Fig. 6.1-35 Example of a mux/demux device utilizing a FGB in a Mach-Zehnder interferometer.

References

1. Webb, J. R., and U. L. Österberg. 1995. In *Optoelectronics for Datacommunication*. San Diego: Academic Press.
2. Liu, M. M. K. 1996. Fiber, cable and coupling. In *Principles and applications of optical communications*, eds. R. C. Lasky, U. L. Österberg, and D. P. Stigliani. Chicago: Irwin.
3. Okoshi, T. 1982. *Optical fibers*. New York: Academic Press.
4. Marcuse, D., D. Gloge, and E. A. J. Marcatiti. 1979. *Guiding properties of fibers. Optical fiber telecommunications*. Orlando: Academic Press.
5. Ghatak, A., and K. Thyagarajan. 1989. *Optical electronics*. Cambridge, UK: Cambridge University Press.
6. Pollock, C. R. 1995. *Fundamentals of optoelectronics*. Chicago: Irwin.
7. Marcuse, D. 1977. *Loss analysis of single-mode finger splices. Bell System Tech. J*. 56: 703–718.
8. Cheo, P. K. 1990. *Fiber optics and optoelectronics*. 2nd. ed. Englewood Cliffs, N.J.: Prentice-Hall.
9. Powers, J. P. 1993. *An introduction to fiber optic systems*. Homewood, I11.: Aksen.
10. Neumann, E. G. 1988. *Single-mode fibers: Fundamentals*. Berlin: Springer-Verlag.
11. Nemoto, S., and T. Makimoto. 1979. Analysis of splice loss in single-mode fibers using a Gaussian field approximation. *Opt. Quantum Electron.* 11:447–457.
12. Maurer, R. D. 1973. Glass fibers for optical communication. *Proc. IEEE* 61: 452–62.
13. Chen, C. L. 1996. *Principles and applications of optical communications*. Chicago: Irwin.
14. Hentschel, C. 1989. *Fiber optics handbook*. Boeblingen, Germany: Hewlett Packard.
15. Allard, F. C. 1990. *Fiber optics handbook. For engineers and scientists*. New York: McGraw-Hill.
16. Shibata, N., and T. Edahiro. 1982. *Trans. Inst. Electron. Commun. Eng. Jpn.* 65: 166–172.
17. Hecht, J. 1987. *Understanding fiber optics*. Indianapolis, Ind.: Sams.
18. Murata, H. 1996. *Handbook of optical fibers and cables*. New York: Dekker.
19. Kalish, D., Key, P. L., Kurkjian, C. R., Tanyal, B. K., and Wang, T. T. 1979. Fiber characterization—mechanical. In *Optical fiber telecommunications*, eds. S. E. Miller and C. G. Chynoweth. Boston: Academic Press.
20. Philen, D. L., and W. T. Anderson. 1988. Optical fiber transmission evaluation. In *Optical fiber telecommunications II*, eds. S. E. Miller and I. P. Kaminow. Boston: Academic Press.
21. Gartside, C. H., P. D. Patel, and M. R. Santana. 1988. Optical fiber cables. In *Optical fiber telecommunications II*, eds. S. E. Miller and I. P. Kaminow. Boston: Academic Press.
22. Radcliffe, J., C. Paddock, and R. Lasky. 1995. Integrated circuits, transceiver modules and packaging. In *Optoelectronics for datacommunication*, eds. R. C. Lasky, U. L. Osterberg, and D. P. Stigliani. San Diego: Academic Press.
23. Young, W. C., and D. R. Frey. 1988. Fiber connectors. In *Optical fiber telecommunications II*, eds. S. E. Miller and I. P. Kaminow. Boston: Academic Press.
24. Molex Fiber Optics, Inc. 1996. *Fiber optic product catalog*, No. 1096.
25. Hill, K. O., Y. Fujii, and D. C. Johnson, *et al.* 1978. "Photosensitivity in Optical Fiber Waveguides: Application to Reflection Filter Fabrication," *Applied Physics Letters* **32**(10): 647–649.
26. Bennion, I., J. A. R. Williams, and L. Zhang, *et al.* 1996. "Tutorial Review, UV-Written In-Fiber Bragg Gratings," *Optical and Quantum Electronics* 28: 93–135.
27. Hill, K. O., D. C. Johnson, and F Bilodeau, *et al*. 1987. "Narrow-Bandwidth Optical Waveguide Transmission Filters: A New Design Concept and Applications to Optical Fiber Communications," *Electronics Letters* **23**(9): 465–66.
28. Bilodeau, R. F., K. O. Hill, and B. Malo, *et al*. 1994. "High-Return-Loss Narrowband All-Fiber Bandpass Bragg Transmission Filter," *IEEE Photonics Technology Letters* **6**(1): 80–82.
29. Hill, K. O., F. Bilodeau, and B. Malo, *et al.* 1994. "Chirped In-Fiber Bragg Grating for Compensation of Optical-Fiber Dispersion," *Optics Letters* **19**(17): 1314–1316.
30. Matsuhara, M., and K. O. Hill. 1974. "Optical-Waveguide Band-Rejection Filters: Design," *Applied Optics* **13**(12): 2886–2888.
31. Hill, K. O., B. Malo, and K. A. Vineberg, *et al*. 1990. "Efficient Mode Conversion in Telecommunication Fiber Using Externally Written Gratings," *Electronics Letters* **26**(16): 1270–1272.
32. Vengsarkar, A. M., J. R. Pedrazzani, and J. B. Judkins, *et al*. 1996. "Long-Period Fiber-Grating-Based Gain Equalizers," *Optics Letters* **21**(5): 336–338.
33. Bilodeau, F., D. C. Johnson, and S. Thériault, *et al*. 1995. "An All-Fiber Dense-Wavelength-Division Multiplexer/Demultiplexer Using Photoimprinted Bragg Gratings," *IEEE Photonics Technology Letters* **7**(4): 388–390.
34. Meltz, G., W. W. Morey, and W. H. Glenn. 1989. "Formation of Bragg Gratings in Optical Fibers by a Transverse Holographic Method," *Optics Letters* **14**(15): 823–825.
35. Hill, K. O., B. Malo, and F. Bilodeau, *et al.* 1993. "Bragg Gratings Fabricated in Monomode Photosensitive Optical Fiber by UV Exposure Through a Phase Mask," *Applied Physics Letters* **62**(10): 1035–1037.

Section **Seven**

Video and image processing

This page intentionally left blank

Chapter 7.1

7.1

Introduction to video

Keith Jack

Although there are many variations and implementation techniques, video signals are just a way of transferring visual information from one point to another. The information may be from a VCR, DVD player, a channel on the local broadcast, cable television, or satellite system, the Internet, or one of many other sources.

Invariably, the video information must be transferred from one device to another. It could be from a satellite set-top box or DVD player to a television. Or it could be from one chip to another inside the satellite set-top box or television. Although it seems simple, there are many different requirements, and therefore many different ways of doing it.

Analog vs. digital

Until a few years ago, most video equipment was designed primarily for analog video. Digital video was confined to professional applications, such as video editing.

The average consumer now uses digital video every day thanks to continuing falling costs. This trend has led to the development of DVD players and recorders, digital set-top boxes, digital television (DTV), portable video players, and the ability to use the Internet for transferring video data.

Video data

Initially, video contained only gray-scale (also called black-and-white) information.

While color broadcasts were being developed, attempts were made to transmit color video using analog RGB (red, green, blue) data. However, this technique occupied 3× more bandwidth than the current gray-scale solution, so alternate methods were developed that led to using Y, R–Y, and G–Y data to represent color information. A technique was then developed to transmit this Y, R–Y, and G–Y information using one signal, instead of three separate signals, and in the same bandwidth as the original gray-scale video signal. This *composite video signal* is what the NTSC, PAL, and SECAM video standards are still based on today.

Today, even though there are many ways of representing video, they are still all related mathematically to RGB. These variations are discussed in more detail in Chapter 7.2.

S-Video was developed for connecting consumer equipment together (it is not used for broadcast purposes). It is a set of two analog signals, one gray-scale (Y) and one that carries the analog R–Y and B–Y color information in a specific format (also called C or chroma). Once available only for S-VHS, it is now supported on most consumer video products.

Although always used by the professional video market, analog RGB video data has made a temporary comeback for connecting high-end consumer equipment together. Like S-Video, it is not used for broadcast purposes.

A variation of the Y, R–Y, and G–Y video signals, called *YPbPr*, is now commonly used for connecting consumer video products together. Its primary advantage is the ability to transfer high-definition video between consumer products. Some manufacturers incorrectly label the YPbPr connectors YUV, YCbCr, or Y(B–Y)(R–Y).

Digital video

The most common digital signals used are *RGB* and *YCbCr*. RGB is simply the digitized version of the analog

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007 Elsevier Inc. All rights reserved

RGB video signals. YCbCr is basically the digitized version of the analog YPbPr video signals, and is the format used by DVD and digital television.

Best connection method

There is always the question of "what is the best connection method for equipment?" For DVD players and digital cable/satellite/terrestrial set-top boxes, the typical order of decreasing video quality is:

1. HDMI (digital YCbCr)
2. HDMI (digital RGB)
3. Analog YPbPr
4. Analog RGB
5. Analog S-Video
6. Analog Composite

Some will disagree about the order. However, most consumer products do digital video processing in the YCbCr color space. Therefore, using YCbCr as the interconnect for equipment reduces the number of color space conversions required. Color space conversion of digital signals is still preferable to D/A (digital-to-analog) conversion followed by A/D (analog-to-digital) conversion, hence the positioning of HDMI RGB above analog YPbPr.

The computer industry has standardized on analog and digital RGB for connecting to the computer monitor.

Video timing

Although it looks like video is continuous motion, it is actually a series of still images, changing fast enough that it looks like continuous motion, as shown in Figure 7.1-1. This typically occurs 50 or 60 times per second for consumer video, and 70–90 times per second for computer displays. Special timing information, called *vertical sync*, is used to indicate when a new image is starting.

Each still image is also composed of *scan lines*, lines of data that occur sequentially one after another down the display, as shown in Figure 7.1-1. Additional timing information, called *horizontal sync*, is used to indicate when a new scan line is starting.

The vertical and horizontal sync information is usually transferred in one of three ways:

1. Separate horizontal and vertical sync signals
2. Separate composite sync signal
3. Composite sync signal embedded within the video signal

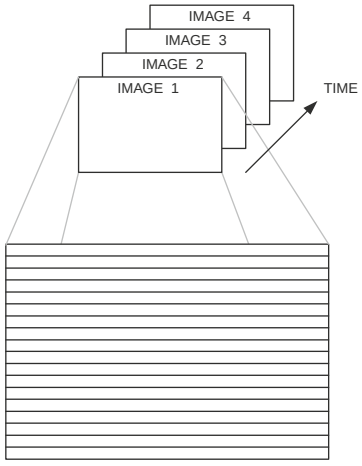


Figure 7.1-1 Video is composed of a series of still images. Each image is composed of individual lines of data.

The composite sync signal is a combination of both vertical and horizontal sync.

Computer and consumer equipment that uses analog RGB video usually uses technique 1 or 2. Consumer equipment that supports composite video or analog YPbPr video usually uses technique 3.

For digital video, either technique 1 is commonly used or timing code words are embedded within the digital video stream.

Interlaced vs. progressive

Since video is a series of still images, it makes sense to simply display each full image consecutively, one after the another.

This is the basic technique of progressive, or non-interlaced, displays. For progressive displays that "paint" an image on the screen, such as a CRT, each image is displayed starting at the top left corner of the display, moving to the right edge of the display. The scanning then moves down one line, and repeats scanning left-to-right. This process is repeated until the entire screen is refreshed, as seen in Figure 7.1-2.

In the early days of television, a technique called "interlacing" was used to reduce the amount of information sent for each image. By transferring the odd-numbered lines, followed by the even-numbered lines (as shown in Figure 7.1-3), the amount of information sent for each image was halved.

Given this advantage of interlacing, why bother to use progressive?

With interlace, each scan line is refreshed half as often as it would be if it were a progressive display. Therefore, to avoid line flicker on sharp edges due to a too-low frame rate, the line-to-line changes are limited, essentially by

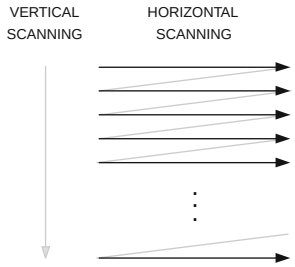


Figure 7.1-2 Progressive displays "paint" the lines of an image consecutively, one after another.

vertically lowpass filtering the image. A progressive display has no limit on the line-to-line changes, so is capable of providing a higher-resolution image (vertically) without flicker.

Today, most broadcasts (including HDTV) are still transmitted as interlaced. Most CRT-based displays are still interlaced while LCD, plasma, and computer displays are progressive.

Video resolution

Video resolution is one of those "fuzzy" things in life. It is common to see video resolutions of 720 × 480 or 1920 × 1080. However, those are just the number of horizontal samples and vertical scan lines, and do not necessarily convey the amount of useful information.

For example, an analog video signal can be sampled at 13.5 MHz to generate 720 samples per line. Sampling the same signal at 27 MHz would generate 1440 samples per line. However, only the number of samples per line has changed, not the resolution of the content.

Therefore, video is usually measured using *lines of resolution*. In essence, how many distinct black and white vertical lines can be seen across the display? This number is then normalized to a 1:1 display aspect ratio (dividing the number by 3/4 for a 4:3 display, or by 9/16 for a 16:9 display). Of course, this results in a lower value for widescreen (16:9) displays, which goes against intuition.

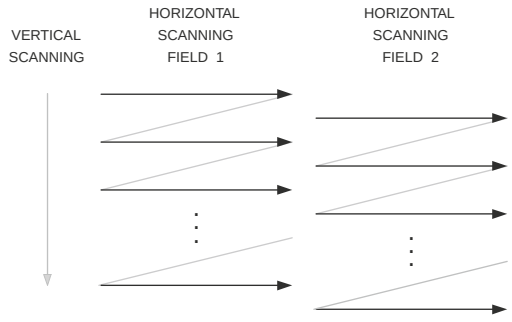


Figure 7.1-3 Interlaced displays "paint" first one-half of the image (odd lines), then the other half (even lines).

Standard-definition

Standard-definition video is usually defined as having 480 or 576 interlaced active scan lines, and is commonly called "480i" and "576i," respectively.

For a fixed-pixel (non-CRT) consumer display with a 4:3 aspect ratio, this translates into an active resolution of 720 × 480i or 720 × 576i. For a 16:9 aspect ratio, this translates into an active resolution of 960 × 480i or 960 × 576i.

Enhanced-definition

Enhanced-definition video is usually defined as having 480 or 576 progressive active scan lines, and is commonly called "480p" and "576p," respectively.

For a fixed-pixel (non-CRT) consumer display with a 4:3 aspect ratio, this translates into an active resolution of 720 × 480p or 720 × 576p. For a 16:9 aspect ratio, this translates into an active resolution of 960 × 480p or 960 × 576p.

The difference between standard and enhanced definition is that standard-definition is interlaced, while enhanced-definition is progressive.

High-definition

High-definition video is usually defined as having 720 progressive (720p) or 1080 interlaced (1080i) active scan lines. For a fixed-pixel (non-CRT) consumer display with a 16:9 aspect ratio, this translates into an active resolution of 1280 × 720p or 1920 × 1080i, respectively.

However, HDTV displays are technically defined as being capable of displaying a minimum of 720p or 1080i active scan lines. They also must be capable of displaying 16:9 content using a minimum of 540 progressive (540p) or 810 interlaced (810i) active scan lines. This enables the manufacturing of CRT-based HDTVs with a 4:3 aspect ratio and LCD/plasma 16:9 aspect ratio displays with resolutions of 1024 ×1024p, 1280 ×768p, 1024 × 768p, and so on, lowering costs.

Audio and video compression

The recent advances in consumer electronics, such as digital television, DVD players and recorders, digital video recorders, and so on, were made possible due to

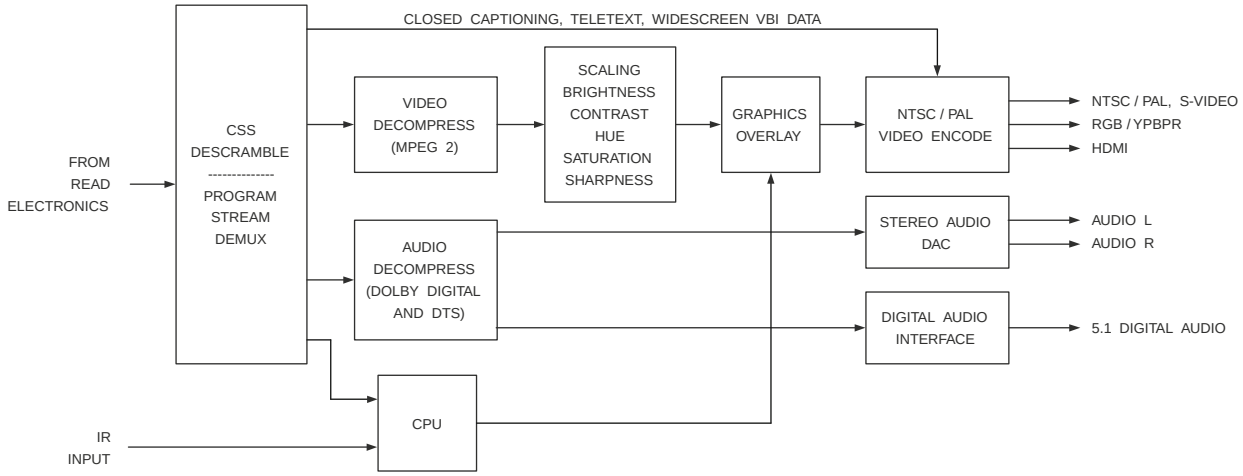


Figure 7.1-4 Simplified block diagram of a basic DVD player.

audio and video compression based largely on MPEG-2 video with Dolby® Digital, DTS®, MPEG-1, or MPEG-2 audio.

New audio and video codecs, such as MPEG-4 HE-AAC, MPEG-4.10 (H.264), and SMPTE 421M (VC-1), offer better compression than previous codecs for the same quality. These advances are enabling new ways of distributing content (both to consumers and within the home), new consumer products (such as portable video players and mobile video/cell phones), and more cable/satellite channels.

Application block diagrams

Looking at a few simplified block diagrams helps envision how video flows through its various operations.

DVD players

Figure 7.1-4 is a simplified block diagram for a basic DVD player, showing the common blocks. Today, all of this is on a single low-cost chip.

In addition to playing DVDs (which are based on MPEG-2 video compression), DVD players are now expected to handle MP3 and WMA audio, MPEG-4 video (for DivX Video), JPEG images, and so on. Special playback modes such as slow/fast forward/reverse at various speeds are also expected. Support for DVD Audio and SACD is also popular.

A recent enhancement to DVD players is the ability to connect to a home network for playing content (music, video, pictures, etc.) residing on the PC. These "networked DVD players" may also include the ability to play movies from the Internet and download content

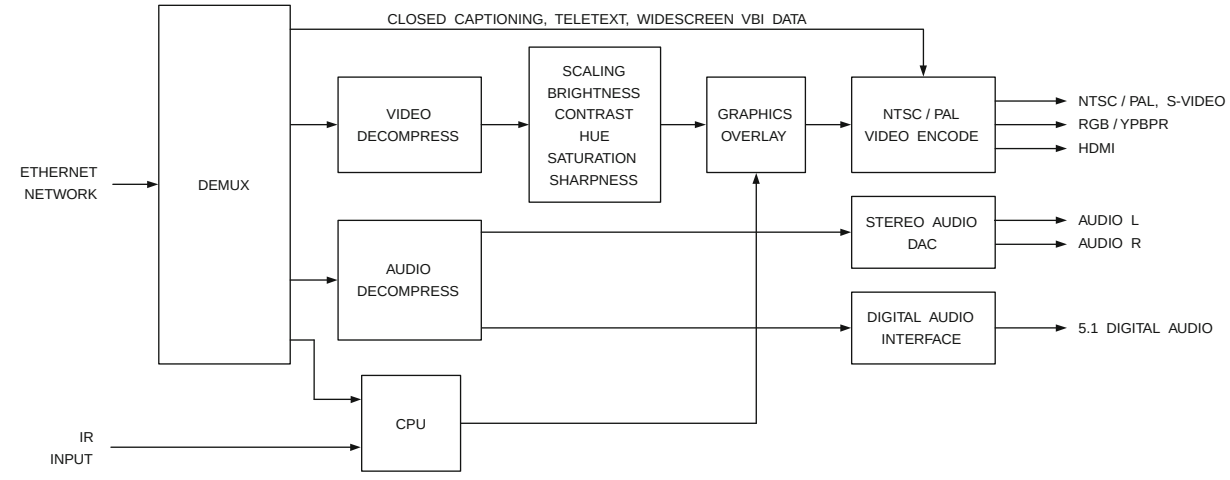


Figure 7.1-5 Simplified block diagram of a digital media adapter.

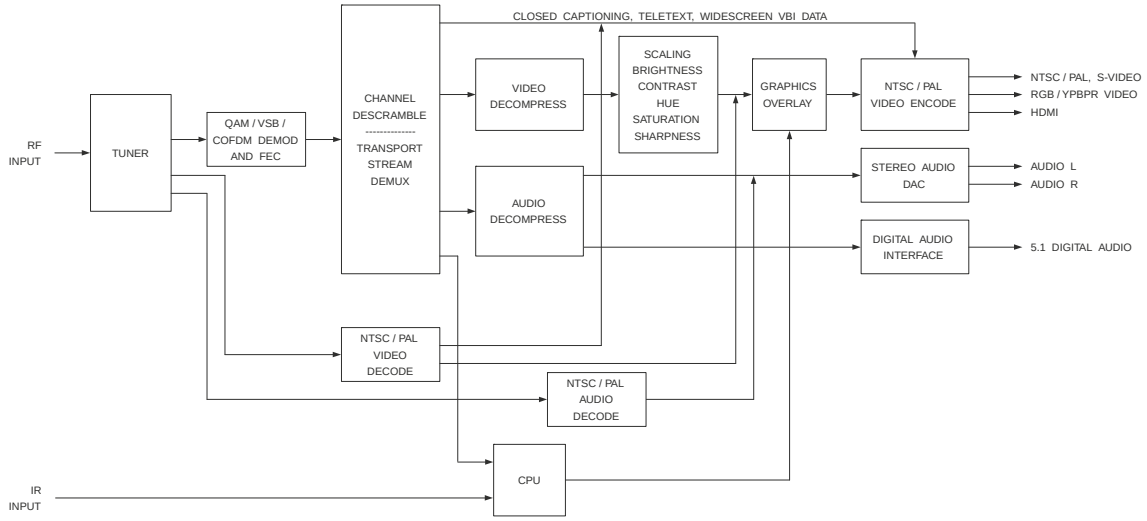


Figure 7.1-6 Simplified block diagram of a digital television set-top box.

onto an internal hard disc drive (HDD) for later viewing. Support for playing audio, video, and pictures from a variety of flash-memory cards is also growing.

In an attempt to look different to quickly grab buyers' attention, some DVD player manufacturers tweak the video frequency response. Since this feature is usually irritating over the long term, it should be defeated or properly adjusted. For the film look many video enthusiasts strive for, the frequency response should be as flat as possible.

Another issue is the output levels of the analog video signals. Although it is easy to generate very accurate video levels, they vary considerably. Reviews now point out this issue since switching between sources may mean changing brightness or black levels, defeating any television calibration or personal adjustments that may have been done by the user.

Digital media adapters

Digital media adapters connect to a home network for playing content (music, video, pictures, and so on) residing on a PC or media server. These small, low-cost boxes enable content to be easily enjoyed on any or all televisions in the home. Many support optional wireless networking, simplifying installation.

Figure 7.1-5 is a simplified block diagram for a basic digital media adapter, showing the common blocks. Today, all of this is on a single low-cost chip.

Digital television set-top boxes

The digital television standards fall into seven major categories:

ATSC (Advanced Television Systems Committee)
DVB (Digital Video Broadcast)
ARIB (Association of Radio Industries and Businesses)
IPTV (including DVB and ARIB over IP)
Open digital cable standards, such as OpenCable
Proprietary digital cable standards
Proprietary digital satellite standards

Originally based on MPEG-2 video and Dolby® Digital or MPEG audio, they now support new advanced audio and video standards, such as MPEG-4 HE-AAC audio, Dolby® Digital Plus audio, MPEG-4.10 (H.264) video, and SMPTE 421M (VC-1) video.

Figure 7.1-6 is a simplified block diagram for a digital television set-top box, showing the common audio and video processing blocks. It is used to receive digital television broadcasts, from either terrestrial (over-the-air), cable, or satellite. A digital television may include this circuitry inside the television.

Many set-top boxes now include two tuners and digital video recorder (DVR) capability. This enables recording one program onto an internal HDD while watching another. Two tuners are also common in digital television receivers to support a picture-in-picture (PIP) feature.

This page intentionally left blank

Chapter 7.2

7.2

Colour spaces

Keith Jack

A color space is a mathematical representation of a set of colors. The three most popular color models are RGB (used in computer graphics); YIQ, YUV, or YCbCr (used in video systems); and CMYK (used in color printing). However, none of these color spaces is directly related to the intuitive notions of hue, saturation, and brightness. This resulted in the temporary pursuit of other models, such as HSI and HSV, to simplify programming, processing, and end-user manipulation.

All of the color spaces can be derived from the RGB information supplied by devices such as cameras and scanners.

RGB color space

The red, green, and blue (RGB) color space is widely used for computer graphics and displays. Red, green, and blue are three primary additive colors (individual components are added together to form a desired color) and are represented by a three-dimensional, Cartesian coordinate system (Figure 7.2-1). The indicated diagonal of the cube, with equal amounts of each primary component, represents various gray levels. Table 7.2-1 contains the RGB values for 100% amplitude, 100% saturated color bars, a common video test signal.

The RGB color space is the most prevalent choice for computer graphics because color displays use red, green, and blue to create the desired color. Therefore, the choice of the RGB color space simplifies the architecture and design of the system. Also, a system that is designed using the RGB color space can take advantage of a large number of existing software routines, since this color space has been around for a number of years.

However, RGB is not very efficient when dealing with real-world images. All three RGB components need to be of equal bandwidth to generate any color within the RGB color cube. The result of this is a frame buffer that has the same pixel depth and display resolution for each RGB component. Also, processing an image in the RGB color space is usually not the most efficient method. For example, to modify the intensity or color of a given pixel, the three RGB values must be read from the frame buffer, the intensity or color calculated, the desired modifications performed, and the new RGB values calculated and written back to the frame buffer. If the system had access to an image stored directly in the intensity and color format, some processing steps would be faster.

For these and other reasons, many video standards use luma and two color difference signals. The most common are the YUV, YIQ, and YCbCr color spaces. Although all are related, there are some differences.

sRGB

Due to the many implementations of the RGB color space, the sRGB color space was formalized. The specification for sRGB (IEC 61966–2–1) uses BT.709 chromaticity, D65 reference white, a display gamma of 2.2, and linear RGB (8 bits per color).

sRGB values have a normalized range of 0–1, with 8-bit digital sRGB values having a range of 0–255 for black–white. A version called "Studio RGB" uses an 8-bit range of 16–235 for black-white, enabling compatibility with video applications.

One limitation of sRGB is that since the normalized values are restricted to the 0–1 range, colors outside the gamut (the triangle produced by them) cannot be used. For this reason, the extended RGB color space, "scRGB," was developed.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007 Elsevier Inc. All rights reserved

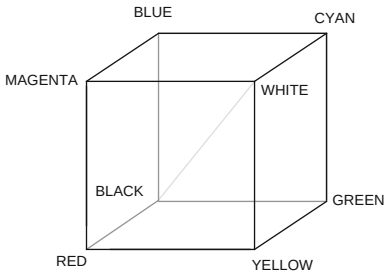


Figure 7.2-1 The RGB Color Cube.

scRGB

The scRGB color space (formerly called sRGB64) extends the dynamic range, color gamut, and bit precision over sRGB. The scRGB gamut is not only much larger than the sRGB gamut, but it is larger than what the human visual system can see. The specification for scRGB (IEC 61966–2–2) uses BT.709 chromaticity, D65 reference white, and linear RGB data (16 bits per color).

Instead of using a normalized range of 0–1, a range of –0.5 to +7.4999 is supported. Values below 0 and above 1 are what enable scRGB to have a larger gamut, compared to sRGB, even though it has the same primary colors. The correlation between the linear 16-bit scRGB values and normalized range are:

$00000 = -0.5$
$04096 = 0.0$ (black)
$12288 = 1.0$ (white)
$16384 = 1.5$
$65535 = 7.4999$

After gamma correction, the correlation between the nonlinear 16-bit scR′G′B′ values and normalized range are:

$00000 = -0.7354$
$04096 = 0.0$ (black)
$12288 = 1.0$ (white)
$65535 = 2.3876$

scRGB to sRGB Conversion

To convert linear 16-bit scRGB to gamma-corrected 8-bit sRGB (notated as $sR'G'B'_8$):

$scR = (scR_{16} / 8192) - 0.5$
$scG = (scG_{16} / 8192) - 0.5$
$scB = (scB_{16} / 8192) - 0.5$

if $(scR_{16}, scG_{16}, scB_{16}) \leq 4095$
$sR'_8 = 0$
$sG'_8 = 0$
$sB'_8 = 0$

if $4096 \leq (scR_{16}, scG_{16}, scB_{16}) \leq 4243$
$sR'_8 = \text{round}[4.500 \times scR \times 255]$
$sG'_8 = \text{round}[4.500 \times scG \times 255]$
$sB'_8 = \text{round}[4.500 \times scB \times 255]$

if $4244 \leq (scR_{16}, scG_{16}, scB_{16}) \leq 12288$
$sR'_8 = \text{round}[(1.099 \times scR^{0.45} - 0.099) \times 255]$
$sG'_8 = \text{round}[(1.099 \times scG^{0.45} - 0.099) \times 255]$
$sB'_8 = \text{round}[(1.099 \times scB^{0.45} - 0.099) \times 255]$

if $(scR_{16}, scG_{16}, scB_{16}) \geq 12289$
$sR'_8 = 255$
$sG'_8 = 255$
$sB'_8 = 255$

YUV Color space

The YUV color space is used by the PAL (Phase Alternation Line), NTSC (National Television System Committee), and SECAM (Sequentiel Couleur Avec Mémoire or Sequential Color with Memory) composite color video standards. The black-and-white system used only luma (Y) information; color information (U and V) was added in such a way that a black-and-white receiver would still display a normal black-and-white picture. Color receivers decoded the additional color information to display a color picture.

The basic equations to convert between gamma-corrected RGB (notated as R′G′B′) and YUV are:

$Y = 0.299R' + 0.587G' + 0.114B'$
$U = -0.147R' - 0.289G' + 0.436B'$
$= 0.492\,(B' - Y)$
$V = 0.615R' - 0.515G' - 0.100B'$
$= 0.877\,(R' - Y)$
$R' = Y + 1.140V$
$G' = Y - 0.395U - 0.581V$
$B' = Y + 2.032U$

Table 7.2-1 100% RGB color bars.

	Nominal range	White	Yellow	Cyan	Green	Magenta	Red	Blue	Black
R	0 to 255	255	255	0	0	255	255	0	0
G	0 to 255	255	255	255	255	0	0	0	0
B	0 to 255	255	0	255	0	255	0	255	0

For digital R′G′B′ values with a range of 0–255, Y has a range of 0–255, U a range of 0 to ±112, and V a range of 0 to ±157. These equations are usually scaled to simplify the implementation in an actual NTSC or PAL digital encoder or decoder.

Note that for digital data, 8-bit YUV and R′G′B′ data should be saturated at the 0 and 255 levels to avoid underflow and overflow wrap-around problems.

If the full range of (B′ − Y) and (R′ − Y) had been used, the composite NTSC and PAL levels would have exceeded what the (then current) black-and-white television transmitters and receivers were capable of supporting. Experimentation determined that modulated subcarrier excursions of 20% of the luma (Y) signal excursion could be permitted above white and below black. The scaling factors were then selected so that the maximum level of 75% amplitude, 100% saturation yellow and cyan color bars would be at the white level (100 IRE).

YIQ Color space

The YIQ color space is derived from the YUV color space and is optionally used by the NTSC composite color video standard. (The "I" stands for "in-phase" and the "Q" for "quadrature," which is the modulation method used to transmit the color information.) The basic equations to convert between R′G′B′ and YIQ are:

$$\begin{aligned} Y &= 0.299R' + 0.587G' + 0.114B' \\ I &= 0.596R' - 0.275G' - 0.321B' \\ &= V\cos 33^\circ - U\sin 33^\circ \\ &= 0.736(R' - Y) - 0.268(B' - Y) \\ Q &= 0.212R' - 0.523G' + 0.311B' \\ &= V\sin 33^\circ + U\cos 33^\circ \\ &= 0.478(R' - Y) + 0.413\ (B' - Y) \end{aligned}$$

or, using matrix notation:

$$\begin{bmatrix} I \\ Q \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \cos(33) & \sin(33) \\ -\sin(33) & \cos(33) \end{bmatrix} \begin{bmatrix} U \\ V \end{bmatrix}$$

$$\begin{aligned} R' &= Y + 0.956I + 0.621Q \\ G' &= Y - 0.272I - 0.647Q \\ B' &= Y - 1.107I + 1.704Q \end{aligned}$$

For digital R′G′B′ values with a range of 0−255, Y has a range of 0−255, I has a range of 0 to ±152, and Q has a range of 0 to ±134. I and Q are obtained by rotating the U and V axes 33°. These equations are usually scaled to simplify the implementation in an actual NTSC digital encoder or decoder.

Note that for digital data, 8-bit YIQ and R′G′B′ data should be saturated at the 0 and 255 levels to avoid underflow and overflow wrap-around problems.

YCbCr Color space

The YCbCr color space was developed as part of ITU-R BT.601 during the development of a world-wide digital component video standard (discussed in Chapter 7.3). YCbCr is a scaled and offset version of the YUV color space. Y is defined to have a nominal 8-bit range of 16−235; Cb and Cr are defined to have a nominal range of 16−240. There are several YCbCr sampling formats, such as 4:4:4, 4:2:2, 4:1:1, and 4:2:0 that are also described.

RGB-YCbCr equations: SDTV

RGB to YCbCr: Analog equations

Many specifications assume the source is analog R′G′B′ with a normalized range of 0−1. This is first converted to analog YPbPr:

$$\begin{aligned} Y &= 0.299R' + 0.587G' + 0.114B' \\ Pb &= -0.169R' - 0.331G' + 0.500B' \\ Pr &= 0.500R' - 0.419G' - 0.081B' \end{aligned}$$

To generate 8-bit YCbCr with the proper values, YPbPr is then quantized to 8 bits:

$$\begin{aligned} Y &= \text{round}\ [219Y + 16] \\ Cb &= \text{round}\ [224Pb + 128] \\ Cr &= \text{round}\ [224Pr + 128] \end{aligned}$$

RGB to YCbCr: Digital Equations

To convert 8-bit digital R′G′B′ data with a 16–235 nominal range (Studio R′G′B′) to YCbCr, the analog equations may be simplified to:

$$\begin{aligned} Y &= 0.299R' + 0.587G' + 0.114B' \\ Cb &= -0.172R' - 0.339G' + 0.511B' + 128 \\ Cr &= 0.511R' - 0.428G' - 0.083B' + 128 \end{aligned}$$

YCbCr to RGB: Analog Equations

Many specifications assume the source is analog YPbPr. This is first converted to analog R′G′B′:

$$\begin{aligned} R' &= Y + 1.402Pr \\ G' &= Y - 0.714Pr - 0.344Pb \\ B' &= Y + 1.772Pb \end{aligned}$$

To generate 8-bit R′G′B′ with a 16−235 nominal range (Studio R′G′B′), R′G′B′ is then quantized to 8 bits:

$$out' = \text{round}[219in' + 16]$$

YCbCr to RGB: Digital Equations

To convert 8-bit YCbCr to R′G′B′ data with a 16–235 nominal range (Studio R′G′B′), the analog equations may be simplified to:

$$R' = Y + 1.371\ (Cr - 128)$$
$$G' = Y - 0.698(Cr - 128) - 0.336(Cb - 128)$$
$$B' = Y + 1.732(Cb - 128)$$

YCbCr to RGB: General Considerations

When performing YCbCr to R′G′B′ conversion, the resulting R′G′B′ values have a nominal range of 16–235, with possible occasional excursions into the 0–15 and 236–255 values. This is due to Y and CbCr occasionally going outside the 16–235 and 16–240 ranges, respectively, due to video processing and noise. Note that 8-bit YCbCr and R′G′B′ data should be saturated at the 0 and 255 levels to avoid underflow and overflow wrap-around problems.

Table 7.2-2 lists the YCbCr values for 75% amplitude, 100% saturated color bars, a common video test signal.

Computer Systems Considerations

If the R′G′B′ data has a range of 0–255, as is commonly found in computer systems, the following equations may be more convenient to use:

$$Y = 0.257R' + 0.504G' + 0.098B' + 16$$
$$Cb = -0.148R' - 0.291G' + 0.439B' + 128$$
$$Cr = 0.439R' - 0.368G' - 0.071B' + 128$$

$$R' = 1.164(Y - 16) + 1.596(Cr - 128)$$
$$G' = 1.164(Y - 16) - 0.813(Cr - 128) - 0.391(Cb - 128)$$
$$B' = 1.164(Y - 16) + 2.018(Cb - 128)$$

Note that 8-bit YCbCr and R′G′B′ data should be saturated at the 0 and 255 levels to avoid underflow and overflow wrap-around problems.

RGB-YCbCr Equations: HDTV

RGB to YCbCr: Analog Equations

Many specifications assume the source is analog R′G′B′ with a normalized range of 0–1. This is first converted to analog YPbPr:

$$Y = 0.213R' + 0.715G' + 0.072B'$$
$$Pb = -0.115R' - 0.385G' + 0.500B'$$
$$Pr = 0.500R' - 0.454G' - 0.046B'$$

To generate 8-bit YCbCr with the proper values, YPbPr is then quantized to 8 bits:

$$Y = \text{round}\ [219Y + 16]$$
$$Cb = \text{round}\ [224Pb + 128]$$
$$Cr = \text{round}\ [224Pr + 128]$$

RGB to YCbCr: Digital Equations

To convert 8-bit digital R′G′B′ data with a 16–235 nominal range (Studio R′G′B′) to YCbCr, the analog equations may be simplified to:

$$Y = 0.213R' + 0.715G' + 0.072B'$$
$$Cb = -0.117R' - 0.394G' + 0.511B' + 128$$
$$Cr = 0.511R' - 0.464G' - 0.047B' + 128$$

YCbCr to RGB: Analog Equations

Many specifications assume the source is analog YPbPr. This is first converted to analog R′G′B′:

$$R' = Y + 1.575Pr$$
$$G' = Y - 0.468Pr - 0.187Pb$$
$$B' = Y + 1.856Pb$$

Table 7.2-2 75% YCbCr color bars.

	Nominal range	White	Yellow	Cyan	Green	Magenta	Red	Blue	Black
					SDTV				
Y	16 to 235	180	162	131	112	84	65	35	16
Cb	16 to 240	128	44	156	72	184	100	212	128
Cr	16 to 240	128	142	44	58	198	212	114	128
					HDTV				
Y	16 to 235	180	168	145	133	63	51	28	16
Cb	16 to 240	128	44	147	63	193	109	212	128
Cr	16 to 240	128	136	44	52	204	212	120	128

To generate 8-bit R′G′B′ with a 16–235 nominal range (Studio R′G′B′), R′G′B′ is then quantized to 8 bits:

$$out' = round[219in' + 16]$$

YCbCr to RGB: Digital Equations

To convert 8-bit YCbCr to R′G′B′ data with a 16–235 nominal range (Studio R′G′B′), the analog equations may be simplified to:

$$R' = Y + 1.540(Cr - 128)$$
$$G' = Y - 0.459(Cr - 128) - 0.183(Cb - 128)$$
$$B' = Y + 1.816(Cb - 128)$$

YCbCr to RGB: General Considerations

When performing YCbCr to R′G′B′ conversion, the resulting R′G′B′ values have a nominal range of 16−235, with possible occasional excursions into the 0−15 and 236−255 values. This is due to Y and CbCr occasionally going outside the 16−235 and 16−240 ranges, respectively, due to video processing and noise. Note that 8-bit YCbCr and R′G′B′ data should be saturated at the 0 and 255 levels to avoid underflow and overflow wrap-around problems.

Table 7.2-2 lists the YCbCr values for 75% amplitude, 100% saturated color bars, a common video test signal.

Computer systems considerations

If the R′G′B′ data has a range of 0−255, as is commonly found in computer systems, the following equations may be more convenient to use:

$$Y = 0.183R' + 0.614G' + 0.062B' + 16$$
$$Cb = -0.101R' - 0.338G' + 0.439B' + 128$$
$$Cr = 0.439R' - 0.399G' - 0.040B' + 128$$

$$R' = 1.164(Y - 16) + 1.793\ (Cr - 128)$$
$$G' = 1.164(Y - 16) - 0.534(Cr - 128) - 0.213(Cb - 128)$$
$$B' = 1.164(Y - 16) + 2.115(Cb - 128)$$

Note that 8-bit YCbCr and R′G′B′ data should be saturated at the 0 and 255 levels to avoid underflow and overflow wrap-around problems.

4:4:4 YCbCr Format

Figure 7.2-2 illustrates the positioning of YCbCr samples for the 4:4:4 format. Each sample has a Y, a Cb, and a Cr value. Each sample is typically 8 bits (consumer applications) or 10 bits (pro-video applications) per component. Each sample therefore requires 24 bits (or 30 bits for pro-video applications).

4:2:2 YCbCr Format

Figure 7.2-3 illustrates the positioning of YCbCr samples for the 4:2:2 format. For every two horizontal Y samples, there is one Cb and Cr sample. Each sample is typically 8 bits (consumer applications) or 10 bits (pro-video applications) per component. Each sample therefore requires 16 bits (or 20 bits for pro-video applications), usually formatted as shown in Figure 7.2-4.

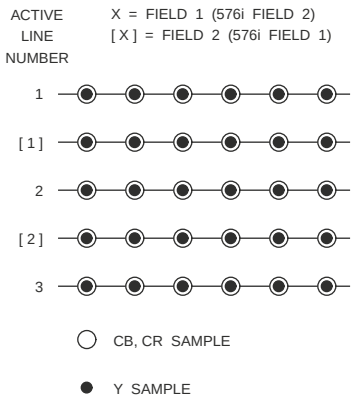


Figure 7.2-2 4:4:4 Co-Sited Sampling. The sampling positions on the active scan lines of an interlaced picture.

To display 4:2:2 YCbCr data, it is first converted to 4:4:4 YCbCr data, using interpolation to generate the missing Cb and Cr samples.

4:1:1 YCbCr Format

Figure 7.2-5 illustrates the positioning of YCbCr samples for the 4:1:1 format (also known as YUV12), used in some consumer video and DV video compression applications. For every four horizontal Y samples, there is one Cb and Cr value. Each component is typically 8 bits. Each sample therefore requires 12 bits, usually formatted as shown in Figure 7.2-6.

To display 4:1:1 YCbCr data, it is first converted to 4:4:4 YCbCr data, using interpolation to generate the missing Cb and Cr samples.

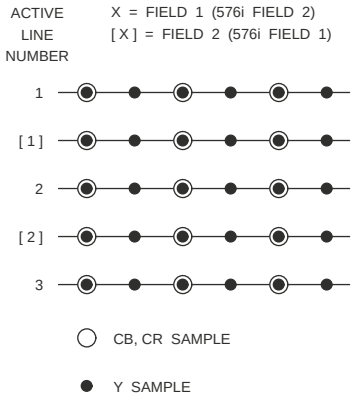


Figure 7.2-3 4:2:2 Co-Sited Sampling. The sampling positions on the active scan lines of an interlaced picture.

SAMPLE 0	SAMPLE 1	SAMPLE 2	SAMPLE 3	SAMPLE 4	SAMPLE 5
Y7 - 0 Y6 - 0 Y5 - 0 Y4 - 0 Y3 - 0 Y2 - 0 Y1 - 0 Y0 - 0	Y7 - 1 Y6 - 1 Y5 - 1 Y4 - 1 Y3 - 1 Y2 - 1 Y1 - 1 Y0 - 1	Y7 - 2 Y6 - 2 Y5 - 2 Y4 - 2 Y3 - 2 Y2 - 2 Y1 - 2 Y0 - 2	Y7 - 3 Y6 - 3 Y5 - 3 Y4 - 3 Y3 - 3 Y2 - 3 Y1 - 3 Y0 - 3	Y7 - 4 Y6 - 4 Y5 - 4 Y4 - 4 Y3 - 4 Y2 - 4 Y1 - 4 Y0 - 4	Y7 - 5 Y6 - 5 Y5 - 5 Y4 - 5 Y3 - 5 Y2 - 5 Y1 - 5 Y0 - 5
CB7 - 0 CB6 - 0 CB5 - 0 CB4 - 0 CB3 - 0 CB2 - 0 CB1 - 0 CB0 - 0	CR7 - 0 CR6 - 0 CR5 - 0 CR4 - 0 CR3 - 0 CR2 - 0 CR1 - 0 CR0 - 0	CB7 - 2 CB6 - 2 CB5 - 2 CB4 - 2 CB3 - 2 CB2 - 2 CB1 - 2 CB0 - 2	CR7 - 2 CR6 - 2 CR5 - 2 CR4 - 2 CR3 - 2 CR2 - 2 CR1 - 2 CR0 - 2	CB7 - 4 CB6 - 4 CB5 - 4 CB4 - 4 CB3 - 4 CB2 - 4 CB1 - 4 CB0 - 4	CR7 - 4 CR6 - 4 CR5 - 4 CR4 - 4 CR3 - 4 CR2 - 4 CR1 - 4 CR0 - 4

16 BITS PER SAMPLE

- 0 = SAMPLE 0 DATA
- 1 = SAMPLE 1 DATA
- 2 = SAMPLE 2 DATA
- 3 = SAMPLE 3 DATA
- 4 = SAMPLE 4 DATA

Figure 7.2-4 4:2:2 Frame Buffer Formatting.

SAMPLE 0	SAMPLE 1	SAMPLE 2	SAMPLE 3	SAMPLE 4	SAMPLE 5
Y7 - 0 Y6 - 0 Y5 - 0 Y4 - 0 Y3 - 0 Y2 - 0 Y1 - 0 Y0 - 0	Y7 - 1 Y6 - 1 Y5 - 1 Y4 - 1 Y3 - 1 Y2 - 1 Y1 - 1 Y0 - 1	Y7 - 2 Y6 - 2 Y5 - 2 Y4 - 2 Y3 - 2 Y2 - 2 Y1 - 2 Y0 - 2	Y7 - 3 Y6 - 3 Y5 - 3 Y4 - 3 Y3 - 3 Y2 - 3 Y1 - 3 Y0 - 3	Y7 - 4 Y6 - 4 Y5 - 4 Y4 - 4 Y3 - 4 Y2 - 4 Y1 - 4 Y0 - 4	Y7 - 5 Y6 - 5 Y5 - 5 Y4 - 5 Y3 - 5 Y2 - 5 Y1 - 5 Y0 - 5
CB7 - 0 CB6 - 0 CR7 - 0 CR6 - 0	CB5 - 0 CB4 - 0 CR5 - 0 CR4 - 0	CB3 - 0 CB2 - 0 CR3 - 0 CR2 - 0	CB1 - 0 CB0 - 0 CR1 - 0 CR0 - 0	CB7 - 4 CB6 - 4 CR7 - 4 CR6 - 4	CB5 - 4 CB4 - 4 CR5 - 4 CR4 - 4

12 BITS PER SAMPLE

- 0 = SAMPLE 0 DATA
- 1 = SAMPLE 1 DATA
- 2 = SAMPLE 2 DATA
- 3 = SAMPLE 3 DATA
- 4 = SAMPLE 4 DATA

Figure 7.2-6 4:1:1 Frame Buffer Formatting.

4:2:0 YCbCr Format

Rather than the horizontal-only 2:1 reduction of Cb and Cr used by 4:2:2, 4:2:0 YCbCr implements a 2:1 reduction of Cb and Cr in both the vertical and horizontal directions. It is commonly used for video compression.

As shown in Figures 7.2-7 through 7.2-11, there are several 4:2:0 sampling formats. Table 7.2-3 lists the YCbCr formats for various DV applications.

To display 4:2:0 YCbCr data, it is first converted to 4:4:4 YCbCr data, using interpolation to generate the new Cb and Cr samples. Note that some solutions do not properly convert the 4:2:0 YCbCr data to the 4:4:4 format, resulting in a "chroma bug."

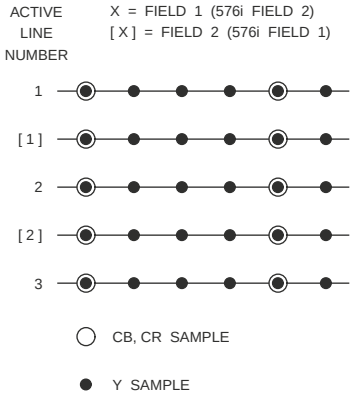


Figure 7.2-5 4:1:1 Co-Sited Sampling. The sampling positions on the active scan lines of an interlaced picture.

xvYCC Color space

The xvYCC (extended gamut YCbCr for video) color space extends the color gamut of normal YCbCr, enabling 1.8× more colors to be reproduced. The specification for xvYCC (IEC 61966−2–4) uses BT.709 chromaticity and D65 reference white. The equations for converting between scR′G′B′ and xvYCbCr are the same as those used for converting between R′G′B′ and YCbCr.

xvYCC-based YCbCr data has an 8-bit range of 1−254, enabling backwards compatibility with existing designs. Y has an 8-bit range of −15/219 to +238/219 (−0.068493 to +1.086758); CbCr has an 8-bit range of −15/224 to +238/224 (−0.066964 to +1.062500). HDMI uses Gamut Boundary Description metadata to convey xvYCC video data is being used.

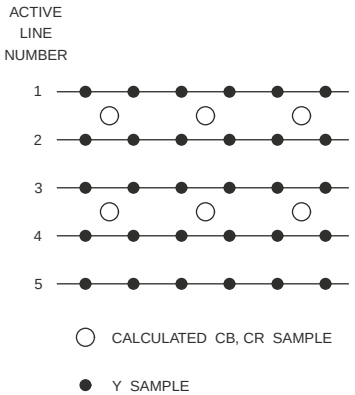


Figure 7.2-7 4:2:0 Sampling for H.261, H.263, and MPEG-1. The sampling positions on the active scan lines of a progressive or noninterlaced picture.

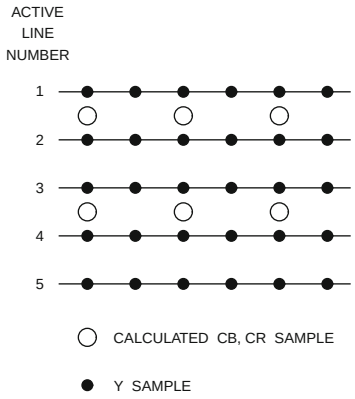


Figure 7.2-8 4:2:0 Sampling for MPEG-2, MPEG-4.2, MPEG-4.10 (H.264), and SMPTE 421M (VC-1). The sampling positions on the active scan lines of a progressive or noninterlaced picture.

PhotoYCC Color space

PhotoYCC (a trademark of Eastman Kodak Company) was developed to encode Photo CD image data. The goal was to develop a display-device-independent color space. For maximum video display efficiency, the color space is based upon ITU-R BT.601 and BT.709.

The encoding process (RGB to PhotoYCC) assumes CIE Standard Illuminant D_{65} and that the spectral sensitivities of the image capture system are proportional to the color-matching functions of the BT.709 reference primaries. The RGB values, unlike those for a computer graphics system, may be negative. PhotoYCC includes colors outside the BT.709 color gamut; these are encoded using negative values.

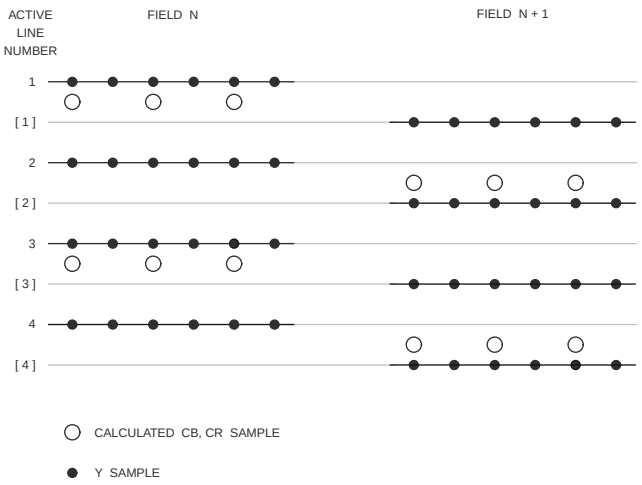


Figure 7.2-9 4:2:0 Sampling for MPEG-2, MPEG-4.2, MPEG-4.10 (H.264), and SMPTE 421M (VC-1). The sampling positions on the active scan lines of an interlaced picture (top_field_first = 1).

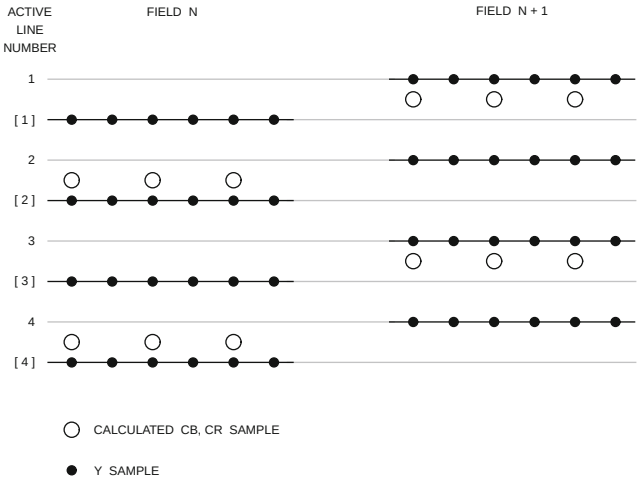


Figure 7.2-10 4:2:0 Sampling for MPEG-2, MPEG-4.2, MPEG-4.10 (H.264), and SMPTE 421M (VC-1). The sampling positions on the active scan lines of an interlaced picture (top_field_first = 0).

RGB to PhotoYCC

Linear RGB data (normalized to have values of 0 to 1) is nonlinearly transformed to PhotoYCC as follows:

for $R, G, B \geq 0.018$

$$R' = 1.099\ R^{0.45} - 0.099$$
$$G' = 1.099\ G^{0.45} - 0.099$$
$$B' = 1.099\ B^{0.45} - 0.099$$

for $-0.018 < R, G, B < 0.018$

$$R' = 4.5\ R$$

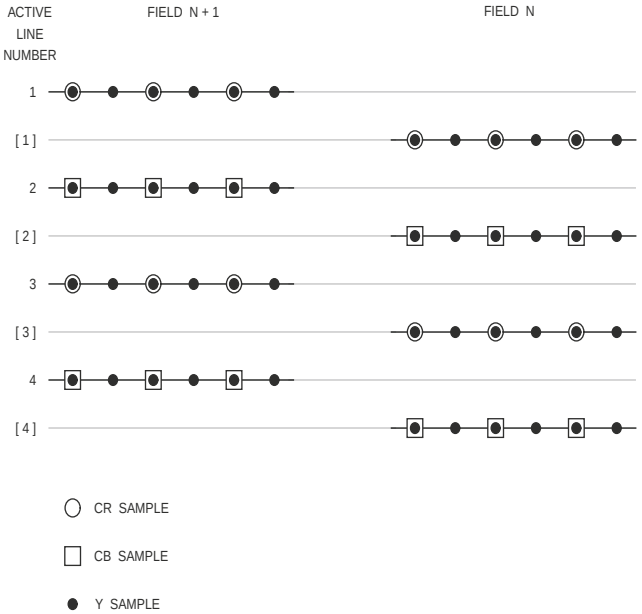


Figure 7.2-11 4:2:0 Co-Sited Sampling for 576i DV and DVCAM. The sampling positions on the active scan lines of an interlaced picture.

Table 7.2-3 YCbCr formats for various DV applications.

	25 Mbps DV					50 Mbps DV			100 Mbps DV				
YCbCr Format	**480-Line DV**	**576-Line DV**	**480-Line DVCAM**	**576-Line DVCAM**	**D-7 (DVCPRO)**	**DVCPRO 50**	**Digital Betacam**	**D-9 (Digital S)**	**DVCPRO HD**	**D-9 HD**	**MPEG-1**	**MPEG-2, -4.2, -4.10 (H.264)**	**H.261, H.263**
4:4:4 Co-Sited												X	
4:2:2 Co-Sited						X	X	X	X	X		X	
4:1:1 Co-Sited	X		X		X								
4:2:0											X	X	X
4:2:0 Co-Sited		X		X									

$G' = 4.5\ G$
$B' = 4.5\ B$

for $R, G, B \leq -0.018$
$R' = -1.099\ |R|^{0.45} - 0.099$
$G' = -1.099\ |G|^{0.45} - 0.099$
$B' = -1.099\ |B|^{0.45} - 0.099$

From R′G′B′ with a 0–255 range, a luma and two chrominance signals (C1 and C2) are generated:

$Y = 0.213R' + 0.419G' + 0.081B'$
$C1 = -0.131R' - 0.256G' + 0.387B' + 156$
$C2 = 0.373R' - 0.312G' - 0.061B' + 137$

As an example, a 20% gray value (R, G, and B = 0.2) would be recorded on the PhotoCD disc using the following values:

$Y = 79$
$C1 = 156$
$C2 = 137$

PhotoYCC to RGB

Since PhotoYCC attempts to preserve the dynamic range of film, decoding PhotoYCC images requires the selection of a color space and range appropriate for the output device. Thus, the decoding equations are not always the exact inverse of the encoding equations. The following equations are suitable for generating RGB values for driving a CRT display, and assume a unity relationship between the luma in the encoded image and the displayed image.

$R' = 0.981Y + 1.315(C2 - 137)$
$G' = 0.981Y - 0.311(C1 - 156) - 0.669(C2 - 137)$
$B' = 0.981Y + 1.601(C1 - 156)$

The R′G′B′ values should be saturated to a range of 0 to 255. The equations above assume the display uses phosphor chromaticities that are the same as the BT.709 reference primaries, and that the video signal luma (V) and the display luminance (L) have the relationship:

for $V \geq 0.0812$
$L = ((V + 0.099) / 1.099)^{1/0.45}$
for $V < 0.0812$
$L = V / 4.5$

HSI, HLS, and HSV Color spaces

The HSI (hue, saturation, intensity) and HSV (hue, saturation, value) color spaces were developed to be more "intuitive" in manipulating color and were designed to approximate the way humans perceive and interpret color. They were developed when colors had to be specified manually, and are rarely used now that users can select colors visually or specify Pantone colors. These color spaces are discussed for historical interest. HLS (hue, lightness, saturation) is similar to HSI; the term lightness is used rather than intensity.

The difference between HSI and HSV is the computation of the brightness component (I or V), which determines the distribution and dynamic range of both the brightness (I or V) and saturation (S). The HSI color space is best for traditional image processing functions such as convolution, equalization, histograms, and so on, which operate by manipulation of the brightness values since I is equally dependent on R, G, and B. The HSV color space is preferred for manipulation of hue and saturation (to shift colors or adjust the amount of color) since it yields a greater dynamic range of saturation.

Figure 7.2-12 illustrates the single hexcone HSV color model. The top of the hexcone corresponds to V = 1, or the maximum intensity colors. The point at the base of the hexcone is black and here V = 0. Complementary colors are 180° opposite one another as measured by H, the angle around the vertical axis (V), with red at 0°. The value of S is a ratio, ranging from 0 on the center line vertical axis (V) to 1 on the sides of the hexcone. Any value of S between 0 and 1 may be associated with the point V = 0. The point S = 0, V = 1 is white. Intermediate values of V for S = 0 are the grays. Note that when S = 0, the value of H is irrelevant. From an artist's viewpoint, any color with V = 1, S = 1 is a pure pigment (whose color is defined by H). Adding white corresponds to decreasing S (without changing V); adding black corresponds to decreasing V (without changing S). Tones are created by decreasing both S and V. Table 7.2-4 lists the 75% amplitude, 100% saturated HSV color bars.

Figure 7.2-13 illustrates the double hexcone HSI color model. The top of the hexcone corresponds to I = 1, or white. The point at the base of the hexcone is black and here

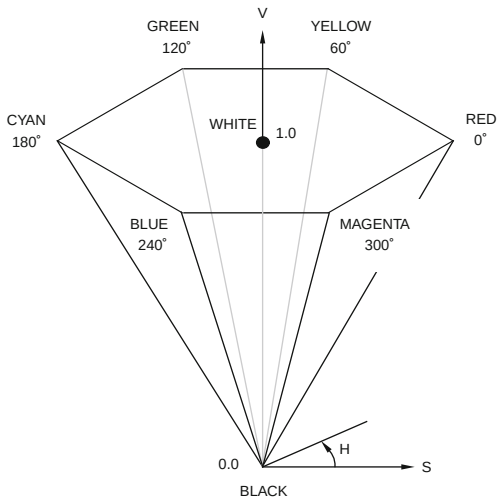


Figure 7.2-12 Single Hexcone HSV Color Model.

Table 7.2-4 75% HSV color bars.

	Nominal range	White	Yellow	Cyan	Green	Magenta	Red	Blue	Black
H	0° to 360°	–	60°	180°	120°	300°	0°	240°	–
S	0 to 1	0	1	1	1	1	1	1	0
V	0 to 1	0.75	0.75	0.75	0.75	0.75	0.75	0.75	0

I = 0. Complementary colors are 180° opposite one another as measured by H, the angle around the vertical axis (I), with red at 0° (for consistency with the HSV model, we have changed from the Tektronix convention of blue at 0°). The value of S ranges from 0 on the vertical axis (I) to 1 on the surfaces of the hexcone. The grays all have S = 0, but maximum saturation of hues is at S = 1, I = 0.5. Table 7.2-5 lists the 75% amplitude, 100% saturated HSI color bars.

Chromaticity diagram

The color gamut perceived by a person with normal vision (the 1931 CIE Standard Observer) is shown in Figure 7.2-14. The diagram and underlying mathematics were updated in 1960 and 1976; however, the NTSC television system is based on the 1931 specifications.

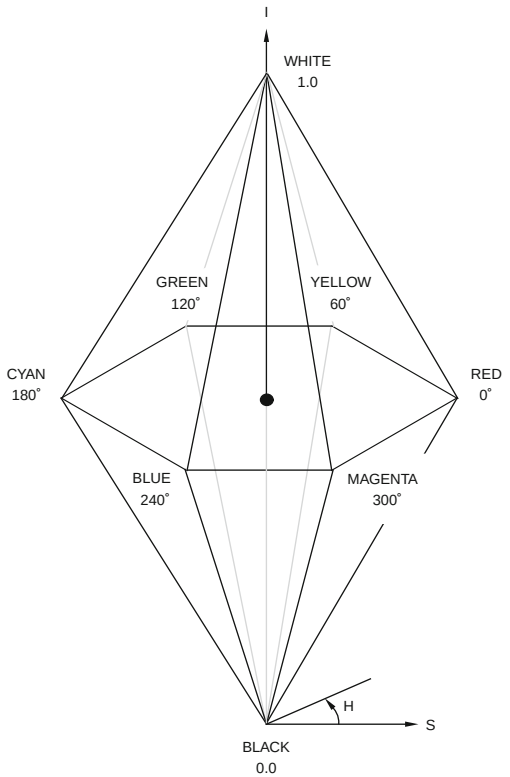


Figure 7.2-13 Double Hexcone HSI Color Model. For consistency with the HSV model, we have changed from the Tektronix convention of blue at 0° and depict the model as a double hexcone rather than as a double cone.

Color perception was measured by viewing combinations of the three standard CIE (International Commission on Illumination or Commission Internationale de I'Eclairage) primary colors: red with a 700-nm wavelength, green at 546.1 nm, and blue at 435.8 nm. These primary colors, and the other spectrally pure colors resulting from mixing of the primary colors, are located along the curved outer boundary line (called the *spectrum locus*), shown in Figure 7.2-14.

The ends of the spectrum locus (at red and blue) are connected by a straight line that represents the purples, which are combinations of red and blue. The area within this closed boundary contains all the colors that can be generated by mixing light of different colors. The closer a color is to the boundary, the more saturated it is. Colors within the boundary are perceived as becoming more pastel as the center of the diagram (white) is approached. Each point on the diagram, representing a unique color, may be identified by its x and y coordinates.

In the CIE system, the intensities of red, green, and blue are transformed into what are called the *tristimulus values*, which are represented by the capital letters X, Y, and Z. These values represent the relative quantities of the primary colors.

The coordinate axes of Figure 7.2-14 are derived from the tristimulus values:

$$\begin{aligned} x &= X/(X+Y+Z) \\ &= \text{red}/(\text{red}+\text{green}+\text{blue}) \end{aligned}$$

$$\begin{aligned} y &= Y/(X+Y+Z) \\ &= \text{green}/(\text{red}+\text{green}+\text{blue}) \end{aligned}$$

$$\begin{aligned} z &= Z/(X+Y+Z) \\ &= \text{blue}/(\text{red}+\text{green}+\text{blue}) \end{aligned}$$

The coordinates x, y, and z are called *chromaticity coordinates*, and they always add up to 1. As a result, z can always be expressed in terms of x and y, which means that only x and y are required to specify any color, and the diagram can be two-dimensional.

Table 7.2-5 75% HSI color bars. For consistency with the HSV model, we have changed from the Tektronix convention of blue at 0°.

	Nominal range	White	Yellow	Cyan	Green	Magenta	Red	Blue	Black
H	0° to 360°	-	60°	180°	120°	300°	0°	240°	-
S	0 to 1	0	1	1	1	1	1	1	0
I	0 to 1	0.375	0.375	0.375	0.375	0.375	0.375	0.375	0

Typically, a source or display specifies three (x, y) coordinates to define the three primary colors it uses. The triangle formed by the three (x, y) coordinates encloses the gamut of colors that the source or display can reproduce. This is shown in Figure 7.2-15, which compares the color gamuts of NTSC, PAL and HDTV. Note that no set of three colors can generate all possible colors, which is why television pictures are never completely accurate.

In addition, a source or display usually specifies the (x, y) coordinate of the white color used, since pure white is not usually captured or reproduced. White is defined as the color captured or produced when all three primary signals are equal, and it has a subtle shade of color to it. Note that luminance, or brightness, information is not included in the standard CIE 1931 chromaticity diagram, but is an axis that is orthogonal to the (x, y) plane. The lighter a color is, the more restricted the chromaticity range is.

The RGB chromaticities and reference white (CIE illuminate C) for the 1953 NTSC standard are:

$$\begin{aligned} R: &\quad x_r = 0.67 \quad y_r = 0.33 \\ G: &\quad x_g = 0.21 \quad y_g = 0.71 \\ B: &\quad x_b = 0.14 \quad y_b = 0.08 \\ \textit{white}: &\quad x_w = 0.3101 \quad y_w = 0.3162 \end{aligned}$$

Modern NTSC, 480i and 480p video systems use a different set of RGB chromaticities (SMPTE "C") and reference white (CIE illuminate D_{65}):

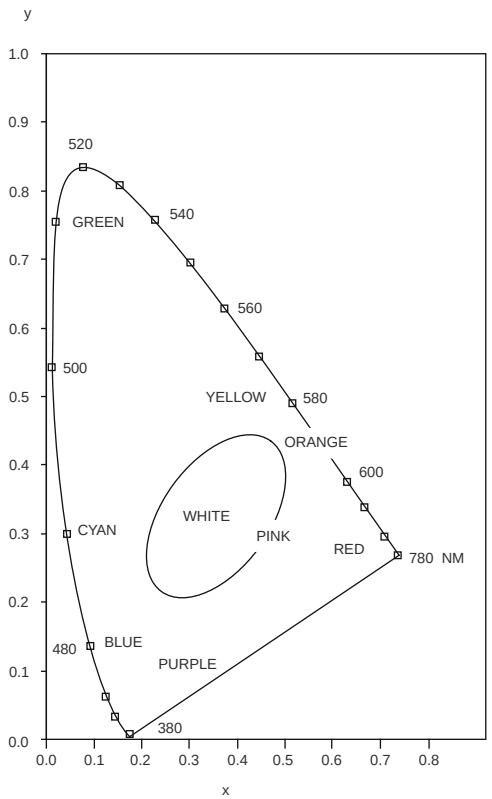


Figure 7.2-14 CIE 1931 Chromaticity Diagram Showing Various Color Regions.

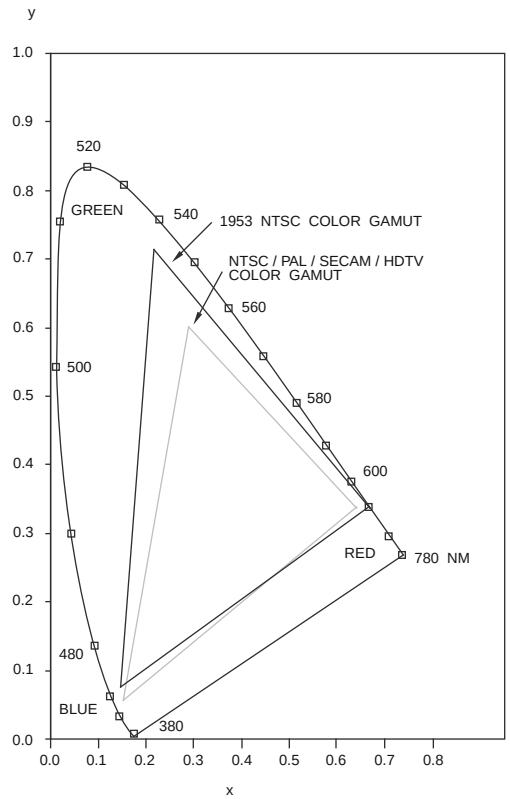


Figure 7.2-15 CIE 1931 Chromaticity Diagram Showing Various Color Gamuts.

R : $x_r = 0.630$ $y_r = 0.340$
G : $x_g = 0.310$ $y_g = 0.595$
B : $x_b = 0.155$ $y_b = 0.070$
white : $x_w = 0.3127$ $y_w = 0.3290$

The RGB chromaticities and reference white (CIE illuminate D_{65}) for PAL, SECAM, 576i and 576p video systems are:

R : $x_r = 0.64$ $y_r = 0.33$
G : $x_g = 0.29$ $y_g = 0.60$
B : $x_b = 0.15$ $y_b = 0.06$
white : $x_w = 0.3127$ $y_w = 0.3290$

The RGB chromaticities and reference white (CIE illuminate D_{65}) for sRGB, scRGB, xvYCC, and HDTV are based on BT.709 and SMPTE 274M:

R : $x_r = 0.64$ $y_r = 0.33$
G : $x_g = 0.30$ $y_g = 0.60$
B : $x_b = 0.15$ $y_b = 0.06$
white : $x_w = 0.3127$ $y_w = 0.3290$

Since different chromaticity and reference white values are used for various video standards, minor color errors may occur when the source and display values do not match; for example, displaying a 480i or 480p program on an HDTV, or displaying an HDTV program on a NTSC television. These minor color errors can easily be corrected at the display by using a 3 × 3 matrix multiplier.

The RGB chromaticities for consumer displays are usually slightly different from the standards. As a result, one or more of the RGB colors are slightly off, such as having too much orange in the red, or too much blue in the green. This can usually be compensated by having the display professionally calibrated.

Non-RGB Color space considerations

When processing information in a non-RGB color space (such as YIQ, YUV, or YCbCr), care must be taken that combinations of values are not created that result in the generation of invalid RGB colors. The term *invalid* refers to RGB components outside the normalized RGB limits of (1,1,1).

For example, given that RGB has a normalized value of (1, 1, 1), the resulting YCbCr value is (235,128,128). If Cb and Cr are manipulated to generate a YCbCr value of (235, 64, 73), the corresponding RGB normalized value becomes (0.6, 1.29, 0.56)—note that the green value exceeds the normalized value of 1.

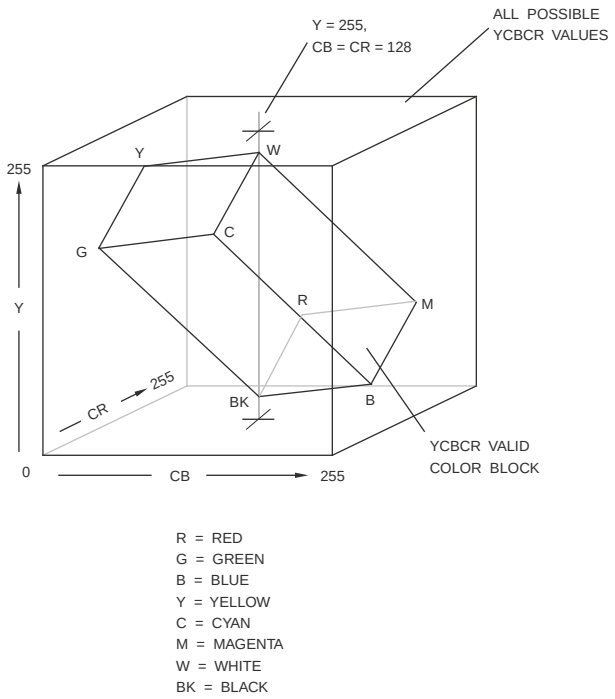


Figure 7.2-16 RGB Limits Transformed into 3-D YCbCr Space.

From this illustration it is obvious that there are many combinations of Y, Cb, and Cr that result in invalid RGB values; these YCbCr values must be processed so as to generate valid RGB values. Figure 7.2-16 shows the RGB normalized limits transformed into the YCbCr color space.

Best results are obtained using a constant luma and constant hue approach—Y is not altered while Cb and Cr are limited to the maximum valid values having the same hue as the invalid color prior to limiting. The constant hue principle corresponds to moving invalid CbCr combinations directly towards the CbCr origin (128, 128), until they lie on the surface of the valid YCbCr color block.

When converting to the RGB color space from a non-RGB color space, care must be taken to include saturation logic to ensure overflow and underflow wrap-around conditions do not occur due to the finite precision of digital circuitry. 8-bit RGB values less than 0 must be set to 0, and values greater than 255 must be set to 255.

Gamma correction

The transfer function of most CRT displays produces an intensity that is proportional to some power (referred to as gamma) of the signal amplitude. As a result, high-intensity ranges are expanded and low-intensity ranges are compressed (see Figure 7.2-17). This is an advantage in combatting noise, as the eye is approximately equally sensitive to equally relative intensity

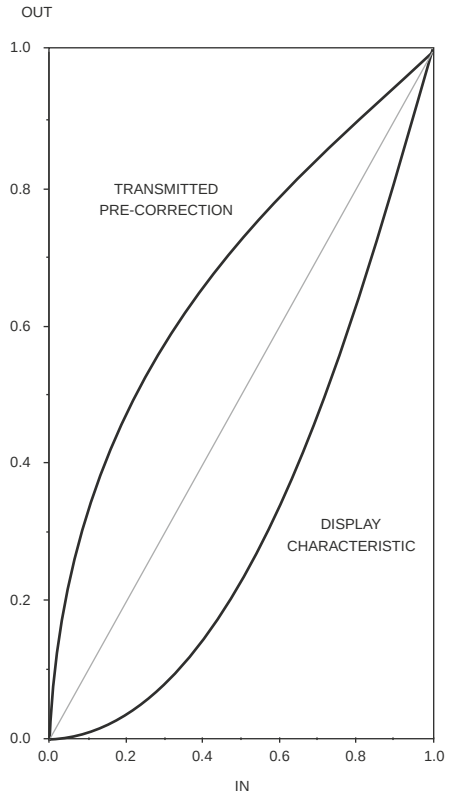


Figure 7.2-17 Effect of Gamma.

changes. By "gamma correcting" the video signals before transmission, the intensity output of the display is roughly linear (the gray line in Figure 7.2-17), and transmission-induced noise is reduced.

To minimize noise in the darker areas of the image, modern video systems limit the gain of the curve in the black region. This technique limits the gain close to black and stretches the remainder of the curve to maintain function and tangent continuity.

Although video standards assume a display gamma of about 2.2, a gamma of about 2.5 is more realistic for CRT displays. However, this difference improves the viewing in a dimly lit environment. More accurate viewing in a brightly lit environment may be accomplished by applying another gamma factor of about 1.14 (2.5/2.2). It is also common to tweak the gamma curve in the display to get closer to the "film look."

Early NTSC systems

Early NTSC systems assumed a simple transform at the display, with a gamma of 2.2. RGB values are normalized to have a range of 0 to1:

$$R = R'^{\,2.2}$$
$$G = G'^{\,2.2}$$
$$B = B'^{\,2.2}$$

To compensate for the nonlinear display, linear RGB data was "gamma-corrected" prior to transmission by the inverse transform. RGB values are normalized to have a range of 0 to 1:

$$R' = R^{1/2.2}$$
$$G' = G^{1/2.2}$$
$$B' = B^{1/2.2}$$

Early PAL and SECAM Systems

Most early PAL and SECAM systems assumed a simple transform at the display, with a gamma of 2.8. RGB values are normalized to have a range of 0 to 1:

$$R = R'^{\,2.8}$$
$$G = G'^{\,2.8}$$
$$B = B'^{\,2.8}$$

To compensate for the nonlinear display, linear RGB data was "gamma-corrected" prior to transmission by the inverse transform. RGB values are normalized to have a range of 0 to 1:

$$R' = R^{1/2.8}$$
$$G' = G^{1/2.8}$$
$$B' = B^{1/2.8}$$

Current systems

Current NTSC, 480i, 480p, and HDTV video systems assume the following transform at the display, with a gamma of [1/0.45]. RGB values are normalized to have a range of 0 to 1:

if $(R', G', B') < 0.081$

$$R = R' / 4.5$$
$$G = G' / 4.5$$
$$B = B' / 4.5$$

if $(R', G', B') \geq 0.081$

$$R = ((R' + 0.099) / 1.099)^{\,1/0.45}$$
$$G = ((G' + 0.099) / 1.099)^{\,1/0.45}$$
$$B = ((B' + 0.099) / 1.099)^{\,1/0.45}$$

Extended gamut color spaces, such as scRGB, do additional processing for below-zero values:

if $(R', G', B') < -0.081$

$$R = -((R' - 0.099) / -1.099)^{\,1/0.45}$$
$$G = -((G' - 0.099) / -1.099)^{\,1/0.45}$$
$$B = -((B' - 0.099) / -1.099)^{\,1/0.45}$$

if $-0.081 \leq (R', G', B') < 0.081$

$$R = R' / 4.5$$
$$G = G' / 4.5$$
$$B = B' / 4.5$$

To compensate for the nonlinear display, linear RGB data is "gamma-corrected" prior to transmission by the inverse transform. RGB values are normalized to have a range of 0 to 1:

if (R, G, B) < 0.018
$R' = 4.5R$
$G' = 4.5G$
$B' = 4.5B$
for (R, G, B) ≥ 0.018
$R' = 1.099R^{0.45} - 0.099$
$G' = 1.099G^{0.45} - 0.099$
$B' = 1.099B^{0.45} - 0.099$

Extended gamut color spaces, such as scRGB, do additional processing for below-zero values:

if (R, G, B) < -0.018
$R' = -1.099\ (-R^{0.45}) + 0.099$
$G' = -1.099\ (-G^{0.45}) + 0.099$
$B' = -1.099\ (-B^{0.45}) + 0.099$
if $-0.018 \leq$ (R, G, B) < 0.018
$R' = 4.5R$
$G' = 4.5G$
$B' = 4.5B$

Although most PAL and SECAM standards specify a gamma of 2.8, a value of [1/0.45] is now commonly used. Thus, these equations are also now used for PAL, SECAM, 576i, and 576p video systems.

Non-CRT displays

Since they are not based on CRTs, the LCD, LCOS, DLP, and plasma displays have different display transforms. To simplify interfacing to these displays, their electronics are designed to accept standard gamma-corrected video and then compensate for the actual transform of the display panel.

Constant luminance problem

Due to the wrong order of the gamma and matrix operations, the U and V (or Cb and Cr) signals also contribute to the luminance (Y) signal. This causes an error in the perceived luminance when the amplitude of U and V is not correct. This may be due to bandwidth-limiting U and V or a non-nominal setting of the U and V gain (color saturation).

For low color frequencies, there is no problem. For high color frequencies, U and V disappear and consequently R′, G′, and B′ degrade to be equal to (only) Y.

References

1. Benson, K. Blair, *Television Engineering Handbook*. McGraw-Hill, Inc., 1986.
2. Devereux, V.G., 1987, *Limiting of YUV digital video signals*, BBC Research Department Report BBC RD1987 22.
3. EIA Standard EIA-189–A, July 1976, *Encoded Color Bar Signal*.
4. Faroudja, Yves Charles, *NTSC and Beyond. IEEE Transactions on Consumer Electronics*, Vol. 34, No. 1, February 1988.
5. IEC 61966–2–1, 1999, *Colour Management–Default RGB Colour Space-sRGB*.
6. IEC 61966–2–2, 2003, *Colour Management–Extended RGB Colour Space-scRGB*.
7. IEC 61966–2–4, 2006, *Colour Management–Extended-Gamut YCC Colour Space for Video Applications–xvYCC*.
8. ITU-R BT.470–6, 1998, *Conventional Television Systems*.
9. ITU-R BT.601–5, 1995, *Studio Encoding Parameters of Digital Television for Standard 4:3 and Widescreen 16:9 Aspect Ratios*.
10. ITU-R BT.709–5, 2002, *Parameter Values for the HDTV Standards for Production and International Programme Exchange*.
11. Photo CD Information Bulletin, *Fully Utilizing Photo CD Images–PhotoYCC Color Encoding and Compression Schemes*, May 1994, Eastman Kodak Company.

Chapter 7.3

7.3

Video signals overview

Keith Jack

Video signals come in a wide variety of options—number of scan lines, interlaced vs. progressive, analog vs. digital, and so on. This chapter provides an overview of the common video signal formats and their timing.

Digital component video background

In digital component video, the video signals are in digital form (YCbCr or R′G′B′), being encoded to composite NTSC, PAL, or SECAM only when it is necessary for broadcasting or recording purposes.

The European Broadcasting Union (EBU) became interested in a standard for digital component video due to the difficulties of exchanging video material between the 576i PAL and SECAM systems. The format held the promise that the digital video signals would be identical whether sourced in a PAL or SECAM country, allowing subsequent encoding to the appropriate composite form for broadcasting. Consultations with the Society of Motion Picture and Television Engineers (SMPTE) resulted in the development of an approach to support international program exchange, including 480i systems.

A series of demonstrations were carried out to determine the quality and suitability for signal processing of various methods. From these investigations, the main parameters of the digital component coding, filtering, and timing were chosen and incorporated into ITU-R BT.601. BT.601 has since served as the starting point for other digital component video standards.

Coding ranges

The selection of the coding ranges balanced the requirements of adequate capacity for signals beyond the normal range and minimizing quantizing distortion. Although the black level of a video signal is reasonably well defined, the white level can be subject to variations due to video signal and equipment tolerances. Noise, gain variations, and transients produced by filtering can produce signal levels outside the nominal ranges.

Either 8 or 10 bits per sample are used for each of the YCbCr or R′G′B′ components. Although 8-bit coding introduces some quantizing distortion, it was originally felt that most video sources contained sufficient noise to mask most of the quantizing distortion. However, if the video source is virtually noise-free, the quantizing distortion is noticeable as contouring in areas where the signal brightness gradually changes. In addition, at least two additional bits of fractional YCbCr or R′G′B′ data were desirable to reduce rounding effects when transmitting between equipment in the studio editing environment. For these reasons, most pro-video equipment uses 10-bit YCbCr or R′G′B′, allowing 2 bits of fractional YCbCr or R′G′B′ data to be maintained.

Initial proposals had equal coding ranges for all three YCbCr components. However, this was changed so that Y had a greater margin for overloads at the white levels, as white level limiting is more visible than black. Thus, the nominal 8-bit Y levels are 16–235, while the nominal 8-bit CbCr levels are 16–240 (with 128 corresponding to no color). Occasional excursions into the other levels are permissible, but never at the 0 and 255 levels.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007 Elsevier Inc. All rights reserved

For 8-bit systems, the values of 0x00 and 0xFF are reserved for timing information. For 10-bit systems, the values of 0x000–0x003 and 0x3FC–0x3FF are reserved for timing information, to maintain compatibility with 8-bit systems.

The YCbCr or R′G′B′ levels to generate 75% color bars are discussed in Chapter 7.2. Digital R′G′B′ signals are defined to have the same nominal levels as Y to provide processing margin and simplify the digital matrix conversions between R′G′B′ and YCbCr.

SDTV sample rate selection

Line-locked sampling of analog R′G′B′ or YUV video signals is done. This technique produces a static orthogonal sampling grid in which samples on the current scan line fall directly beneath those on previous scan lines and fields, as shown Figures 7.2-2 through 7.2-11.

Another important feature is that the sampling is locked in phase so that one sample is coincident with the 50% amplitude point of the falling edge of analog horizontal sync (0×0). This ensures that different sources produce samples at nominally the same positions in the picture. Making this feature common simplifies conversion from one standard to another.

For 480i and 576i video systems, several Y sampling frequencies were initially examined, including four times F_{sc}. However, the four times F_{sc} sampling rates did not support the requirement of simplifying international exchange of programs, so they were dropped in favor of a single common sampling rate. Because the lowest sample rate possible (while still supporting quality video) was a goal, a 12 MHz sample rate was preferred for a long time, but eventually was considered to be too close to the Nyquist limit, complicating the filtering requirements. When the frequencies between 12 MHz and 14.3 MHz were examined, it became evident that a 13.5 MHz sample rate for Y provided some commonality between 480i and 576i systems. Cb and Cr, being color difference signals, do not require the same bandwidth as the Y, so may be sampled at one-half the Y sample rate, or 6.75 MHz.

The "4:2:2" notation now commonly used originally applied to NTSC and PAL video, implying that Y, U and V were sampled at 4×, 2×, and 2× the color subcarrier frequency, respectively. The "4:2:2" notation was then adapted to BT.601 digital component video, implying that the sampling frequencies of Y, Cb and Cr were 4×, 2×, and 2× 3.375 MHz, respectively. "4:2:2" now commonly means that the sample rate of Cb and Cr is one-half that of Y, regardless of the actual sample rates used.

With 13.5 MHz sampling, each scan line contains 858 samples (480i systems) or 864 samples (576i systems) and consists of a digital blanking interval followed by an active line period. Both the 480i and 576i systems use 720 samples during the active line period. Having a common number of samples for the active line period simplifies the design of multistandard equipment and standards conversion. With a sample rate of 6.75 MHz for Cb and Cr (4:2:2 sampling), each active line period contains 360 Cr samples and 360 Cb samples.

With analog systems, problems may arise with repeated processing, causing an extension of the blanking intervals and softening of the blanking edges. Using 720 digital samples for the active line period accommodates the range of analog blanking tolerances of both the 480i and 576i systems. Therefore, repeated processing may be done without affecting the digital blanking interval. Blanking to define the analog picture width need only be done once, preferably at the display or upon conversion to analog video.

Initially, BT.601 supported only 480i and 576i systems with a 4:3 aspect ratio (720 × 480i and 720 × 576i active resolutions). Support for a 16:9 aspect ratio was then added (960 × 480i and 960 × 576i active resolutions) using an 18 MHz sample rate.

EDTV sample rate selection

ITU BT.1358 defines the progressive SDTV video signals, also known as 480p or 576p, or Enhanced Digital Television (EDTV). The sample rate is doubled to 27 MHz (4:3 aspect ratio) or 36 MHz (16:9 aspect ratio) in order to keep the same static orthogonal sampling grid as that used by BT.601.

HDTV sample rate selection

ITU BT.709 defines the 720p, 1080i, and 1080p video signals, respectively. With HDTV, a different technique was used—the number of active samples per line and the number of active lines per frame is constant regardless of the frame rate. Thus, in order to keep a static orthogonal sampling grid, each frame rate uses a different sample clock rate.

480i and 480p systems

Interlaced analog composite video

(M) NTSC and (M) PAL are analog composite video signals that carry all timing and color information within a single signal. These analog interfaces use 525 lines per frame.

Interlaced analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 480i (since there

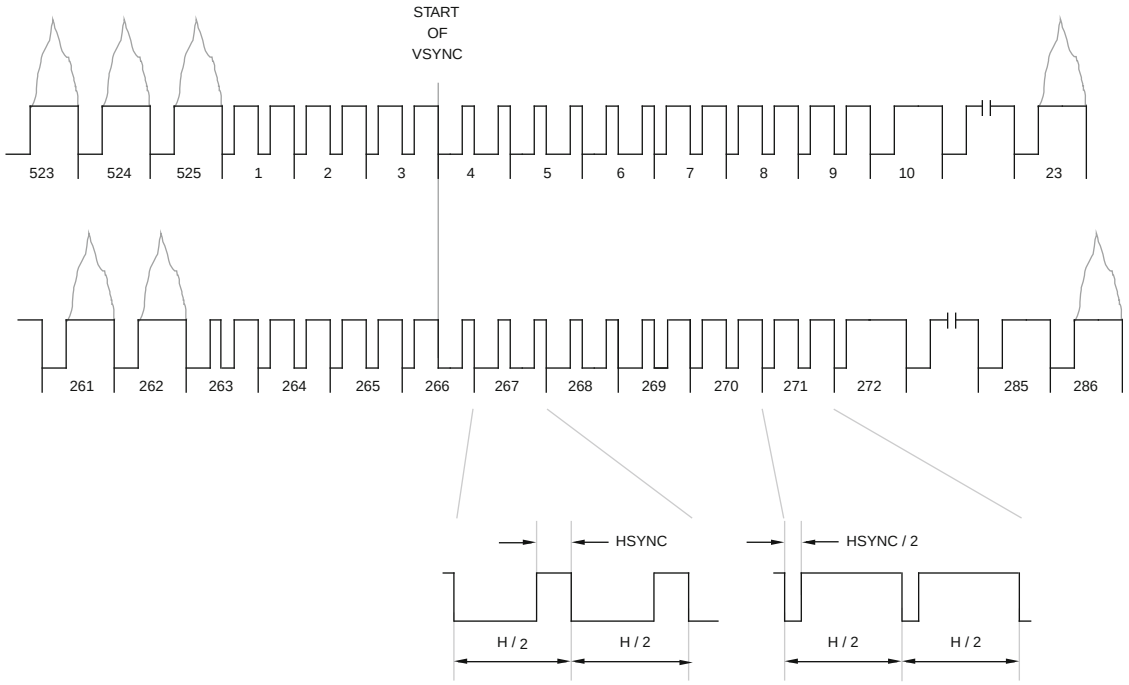


Figure 7.3-1 480i Vertical Interval Timing.



Figure 7.3-2 480p Vertical Interval Timing.

are typically 480 active scan lines per frame and they are interlaced), the frame rate is usually 29.97 Hz (30/1.001) for compatibility with (M) NTSC timing. The analog interface uses 525 lines per frame, with active video present on lines 23–262 and 286–525, as shown in Figure 7.3-1.

For the 29.97 Hz frame rate, each scan line time (H) is about 63.556 µs. Detailed horizontal timing is dependent on the specific video interface used.

Progressive analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 480p (since there are typically 480 active scan lines per frame and they are progressive), the frame rate is usually 59.94 Hz (60/1.001) for easier compatibility with (M) NTSC timing. The analog interface uses 525 lines per frame, with active video present on lines 45–524, as shown in Figure 7.3-2.

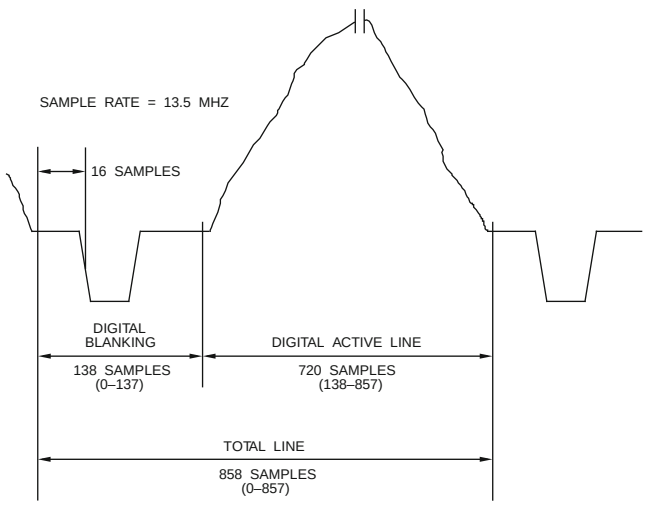


Figure 7.3-3 480i Analog–Digital Relationship (4:3 Aspect Ratio, 29.97 Hz Frame Rate, 13.5 MHz Sample Clock). BT.601 specifies 16 samples for the front porch; CEA-861D (DVI and HDMI timing) specifies 19 samples for the front porch.

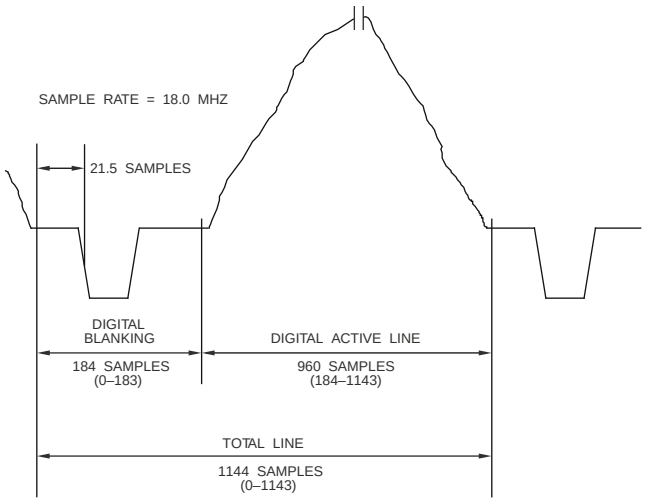


Figure 7.3-4 480i Analog–Digital Relationship (16:9 Aspect Ratio, 29.97 Hz Frame Rate, 18 MHz Sample Clock).

For the 59.94 Hz frame rate, each scan line time (H) is about 31.776 µs. Detailed horizontal timing is dependent on the specific video interface used.

Interlaced digital component video

BT.601 and SMPTE 267M specify the representation for 480i digital R′G′B′ or YCbCr video signals. Active resolutions defined within BT.601 and SMPTE 267M, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

960 × 480i	18.0 MHz	29.97 Hz
720 × 480i	13.5 MHz	29.97 Hz

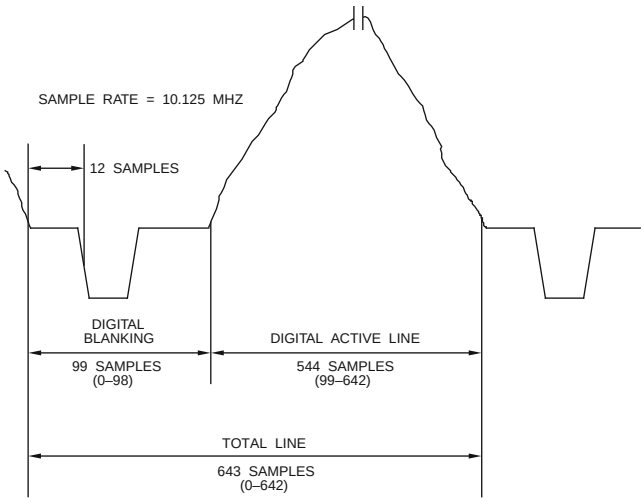


Figure 7.3-6 480i Analog–Digital Relationship (4:3 Aspect Ratio, 29.97 Hz Frame Rate, 10.125 MHz Sample Clock).

Other common active resolutions, their 1× sample rates (F_s), and frame rates, are:

864 × 480i	16.38 MHz	29.97 Hz
704 × 480i	13.50 MHz	29.97 Hz
640 × 480i	12.27 MHz	29.97 Hz
544 × 480i	10.12 MHz	29.97 Hz
528 × 480i	9.900 MHz	29.97 Hz
480 × 480i	9.000 MHz	29.97 Hz
352 × 480i	6.750 MHz	29.97 Hz

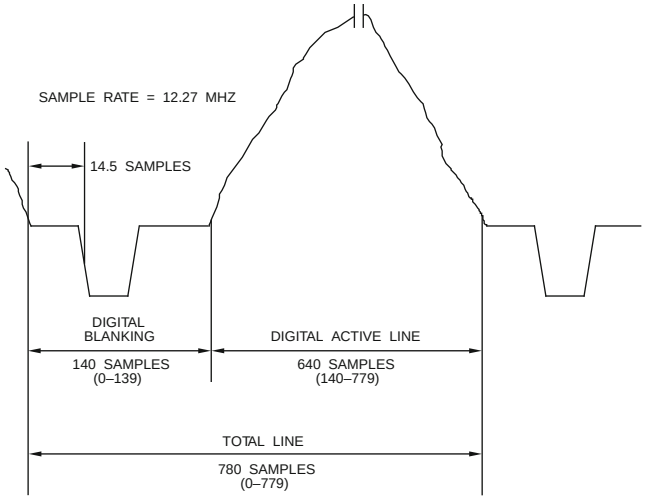


Figure 7.3-5 480i Analog–Digital Relationship (4:3 Aspect Ratio, 29.97 Hz Frame Rate, 12.27 MHz Sample Clock).

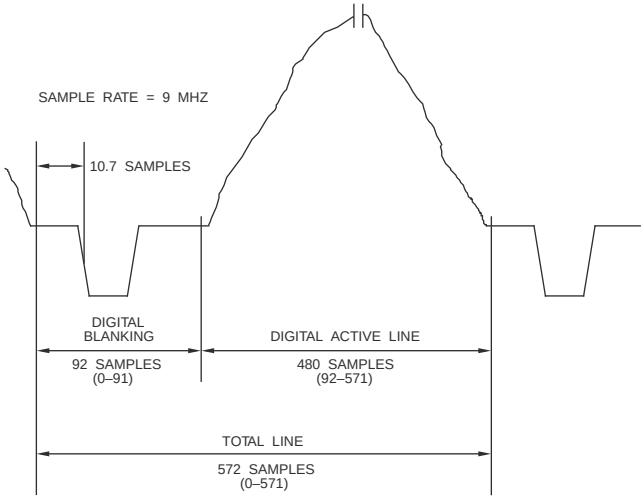
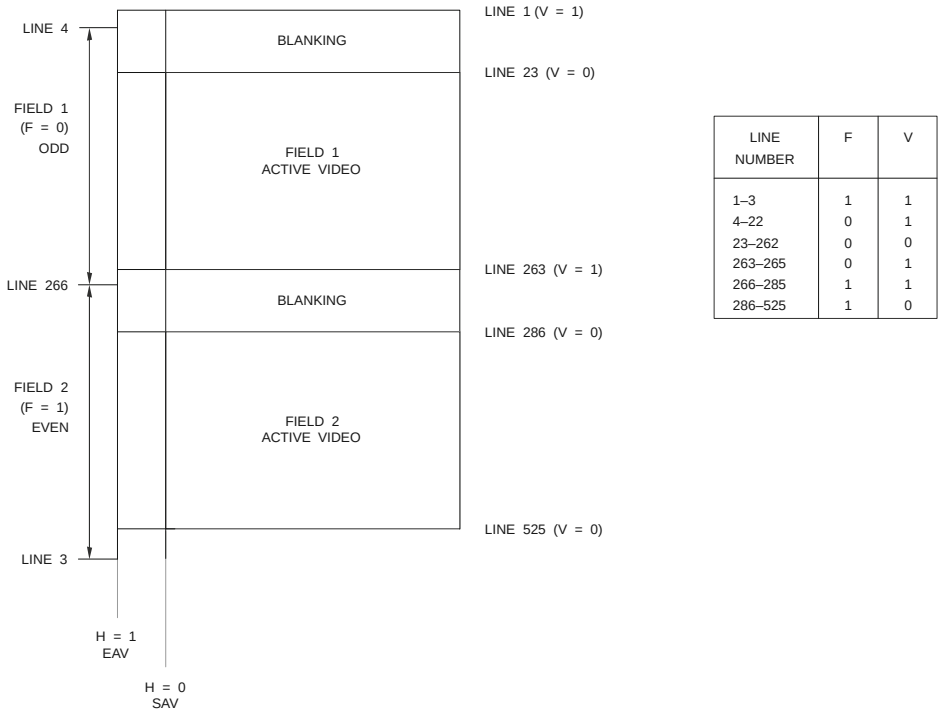


Figure 7.3-7 480i Analog–Digital Relationship (4:3 Aspect Ratio, 29.97 Hz Frame Rate, 9 MHz Sample Clock).



LINE NUMBER	F	V
1–3	1	1
4–22	0	1
23–262	0	0
263–265	0	1
266–285	1	1
286–525	1	0

Figure 7.3-8 480i Digital Vertical Timing (480 Active Lines). F and V change state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-3 through 7.3-7. These active lines are used by the SMPTE RP-202, ATSC A/54a, and ARIB STD-B32 standards. CEA-861D (DVI and HDMI timing) specifies lines 22–261 and 285–524 for active video. IEC 61834-2, ITU-R BT.1618, and SMPTE 314M (DV formats) specify lines 23–262 and 285–524 for active video. ITU-R BT.656 specifies lines 20–263 and 283–525 for active video, resulting in 487 total active lines per frame.

864 × 480i is a 16:9 square pixel format, while 640 × 480i is a 4:3 square pixel format. Although the ideal 16:9 resolution is 854 × 480i, 864 × 480i supports the MPEG 16 × 16 block structure. The 704 × 480i format is done by using the 720 × 480i format, and blanking the first eight and last eight samples each active scan line. Example relationships between the analog and digital signals are shown in Figures 7.3-3 through 7.3-7.

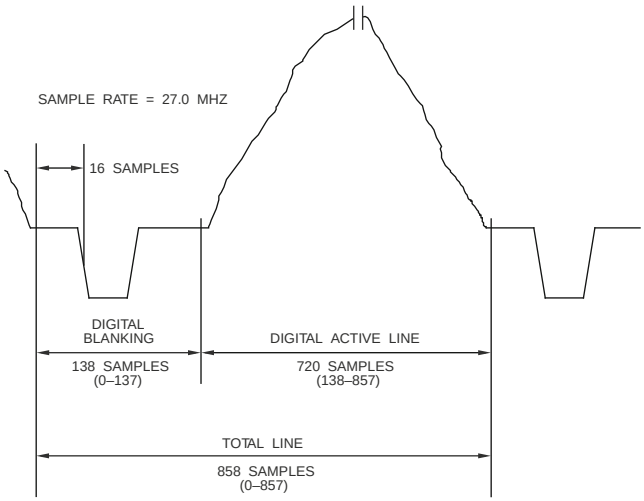


Figure 7.3-9 480p Analog–Digital Relationship (4:3 Aspect Ratio, 59.94 Hz Frame Rate, 27 MHz Sample Clock).

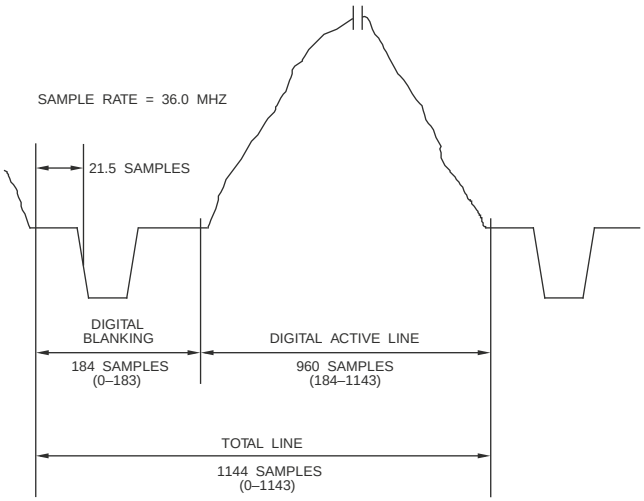


Figure 7.3-10 480p Analog–Digital Relationship (16:9 Aspect Ratio, 59.94 Hz Frame Rate, 36 MHz Sample Clock).

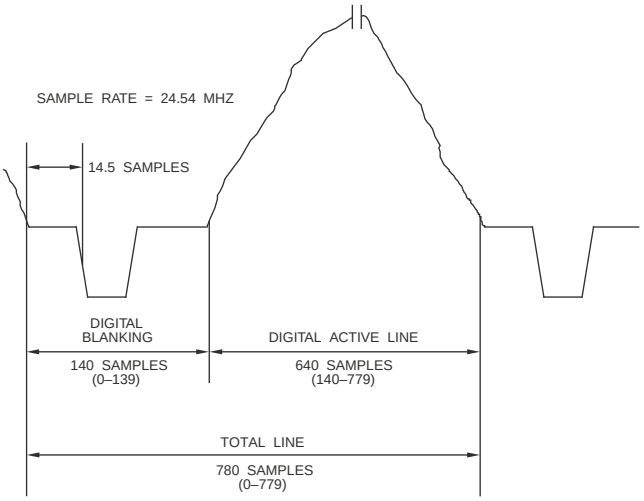


Figure 7.3-11 480p Analog–Digital Relationship (4:3 Aspect Ratio, 59.94 Hz Frame Rate, 24.54 MHz Sample Clock).

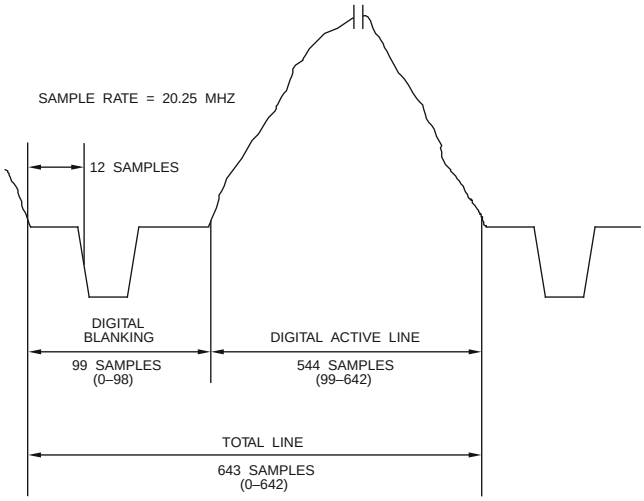


Figure 7.3-12 480p Analog–Digital Relationship (4:3 Aspect Ratio, 59.94 Hz Frame Rate, 20.25 MHz Sample Clock).

The H (horizontal blanking), V (vertical blanking), and F (field) signals are defined in Figure 7.3-8. The H, V, and F timing indicated is compatible with video compression standards rather than BT.656.

Progressive digital component video

BT.1358 and SMPTE 293M specify the representation for 480p digital R′G′B′ or YCbCr video signals. Active resolutions defined within BT.1358 and SMPTE 293M, their 1× sample rates (Fs), and frame rates, are:

960 × 480p	36.0 MHz	59.94 Hz
720 × 480p	27.0 MHz	59.94 Hz

Other common active resolutions, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

864 × 480p	32.75 MHz	59.94 Hz
704 × 480p	27.00 MHz	59.94 Hz
640 × 480p	24.54 MHz	59.94 Hz
544 × 480p	20.25 MHz	59.94 Hz
528 × 480p	19.80 MHz	59.94 Hz
480 × 480p	18.00 MHz	59.94 Hz
352 × 480p	13.50 MHz	59.94 Hz

864 × 480p is a 16:9 square pixel format, while 640 × 480p is a 4:3 square pixel format. Although the ideal 16:9 resolution is 854 × 480p, 864 × 480p supports the MPEG 16 × 16 block structure. The 704 × 480p format is done by using the 720 × 480p format, and blanking the first eight and last eight samples each active scan line. Example relationships between the analog and digital signals are shown in Figures 7.3-9 through 7.3-12.

The H (horizontal blanking), V (vertical blanking), and F (field) signals are defined in Figure 7.3-13. The H, V, and F timing indicated is compatible with video compression standards rather than BT.656.

SIF and QSIF

SIF is defined to have an active resolution of 352 × 240p. Square pixel SIF is defined to have an active resolution of 320 × 240p.

QSIF is defined to have an active resolution of 176 × 120p. Square pixel QSIF is defined to have an active resolution of 160 × 120p.

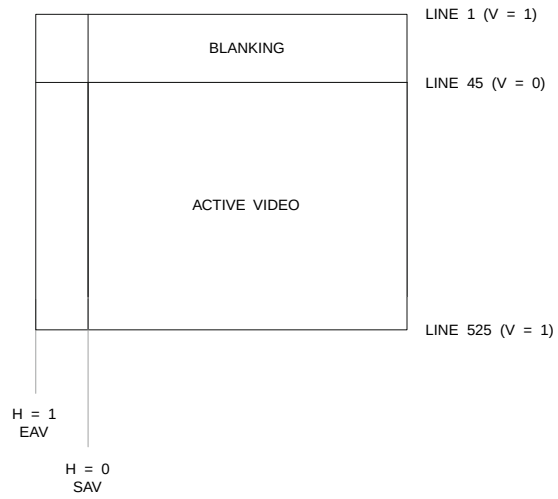
576i and 576p systems

Interlaced analog composite video

(B, D, G, H, I, N, N_C) PAL are analog composite video signals that carry all timing and color information within a single signal. These analog interfaces use 625 lines per frame.

Interlaced analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 576i (since there are typically 576 active scan lines per frame and



LINE NUMBER	F	V
1–44	0	1
45–524	0	0
525	0	1

Figure 7.3-13 480p Digital Vertical Timing (480 Active Lines). V changes state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-9 through 7.3-12. These active lines are used by the SMPTE RP-202, ATSC A/54, and ARIB STD-B32 standards. However, CEA-861 (DVI and HDMI timing) specifies lines 43×522 for active video.

they are interlaced), the frame rate is usually 25 Hz for compatibility with PAL timing. The analog interface uses 625 lines per frame, with active video present on lines 23–310 and 336–623, as shown in Figure 7.3-14.

For the 25 Hz frame rate, each scan line time (H) is 64 µs. Detailed horizontal timing is dependent on the specific video interface used.

Progressive analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 576p (since there are typically 576 active scan lines per frame and they are progressive), the frame rate is usually 50 Hz for compatibility with PAL timing. The analog interface

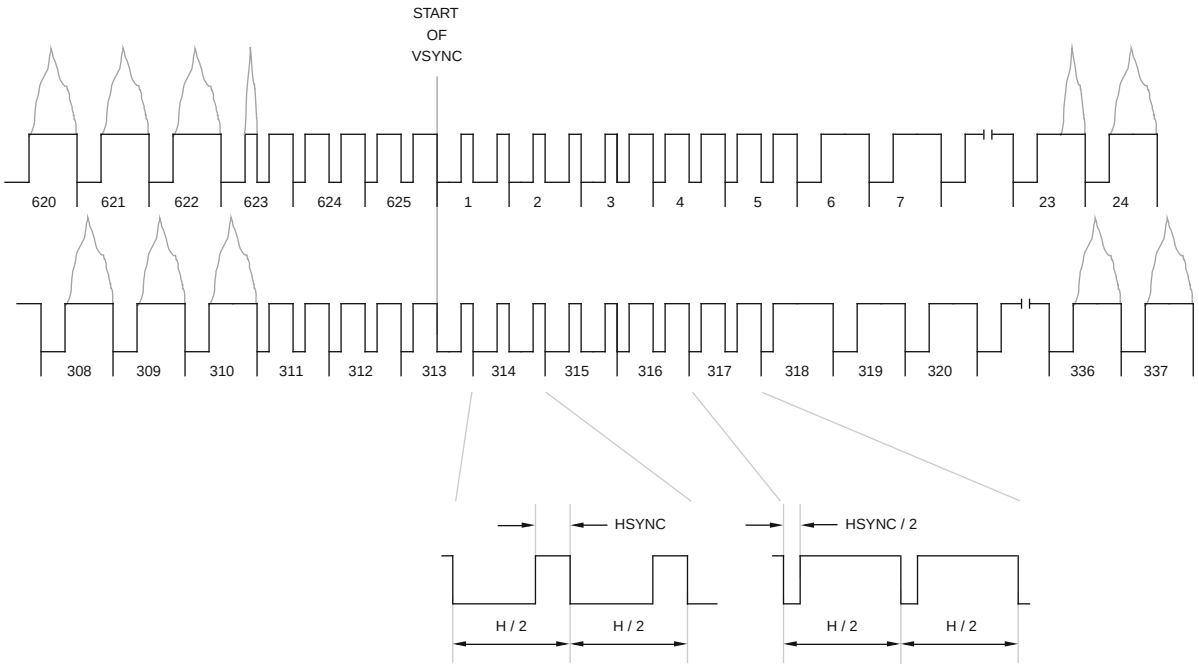


Figure 7.3-14 576i Vertical Interval Timing.

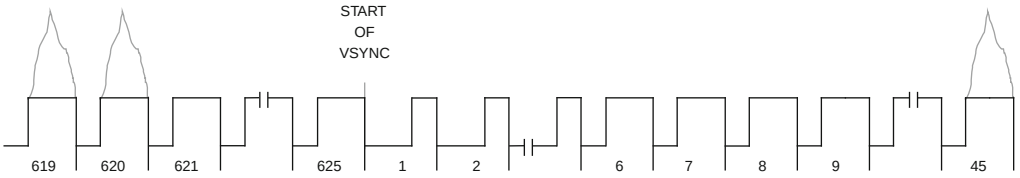


Figure 7.3-15 576p Vertical Interval Timing.

uses 625 lines per frame, with active video present on lines 45–620, as shown in Figure 7.3-15.

For the 50 Hz frame rate, each scan line time (H) is 32 µs. Detailed horizontal timing is dependent on the specific video interface used.

Interlaced digital component video

BT.601 specifies the representation for 576i digital R′G′B′ or YCbCr video signals. Active resolutions defined within BT.601, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

960 × 576i	18.0 MHz	25 Hz
720 × 576i	13.5 MHz	25 Hz

Other common active resolutions, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

1024 × 576i	19.67 MHz	25 Hz
768 × 576i	14.75 MHz	25 Hz
704 × 576i	13.50 MHz	25 Hz
544 × 576i	10.12 MHz	25 Hz
480 × 576i	9.000 MHz	25 Hz

1024 × 576i is a 16:9 square pixel format, while 768 × 576i is a 4:3 square pixel format. The 704 × 576i format is done by using the 720 × 576i format, and blanking the first eight and last eight samples each active scan line. Example relationships between the analog and digital signals are shown in Figures 7.3-16 through 7.3-19.

The H (horizontal blanking), V (vertical blanking), and F (field) signals are defined in Figure 7.3-20. The H, V, and F timing indicated is compatible with video compression standards rather than BT.656.

Progressive digital component video

BT.1358 specifies the representation for 576p digital R′G′B′ or YCbCr signals. Active resolutions defined within BT.1358, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

960 × 576p	36.0 MHz	25 Hz
720 × 576p	27.0 MHz	25 Hz

Other common active resolutions, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

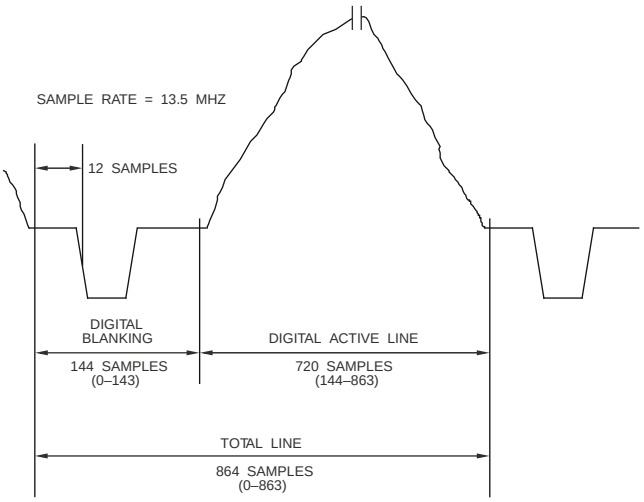


Figure 7.3-16 576i Analog–Digital Relationship (4:3 Aspect Ratio, 25 Hz Frame Rate, 13.5 MHz Sample Clock).

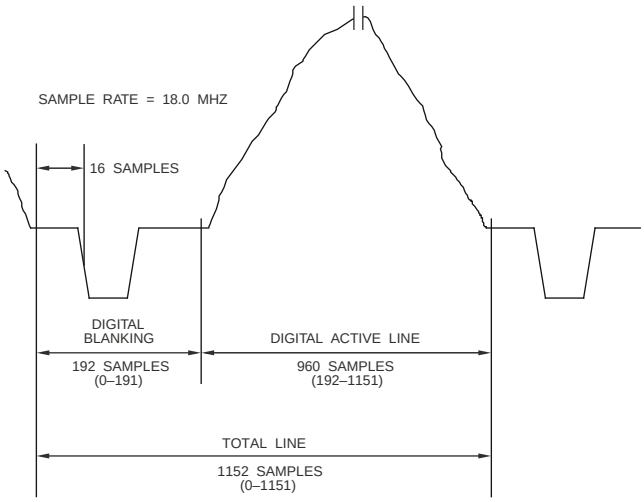


Figure 7.3-17 576i Analog–Digital Relationship (16:9 Aspect Ratio, 25 Hz Frame Rate, 18 MHz Sample Clock).

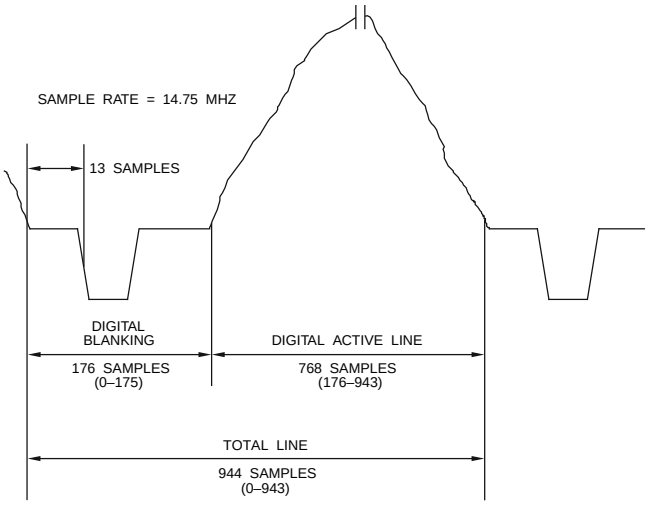


Figure 7.3-18 576i Analog–Digital Relationship (4:3 Aspect Ratio, 25 Hz Frame Rate, 14.75 MHz Sample Clock).

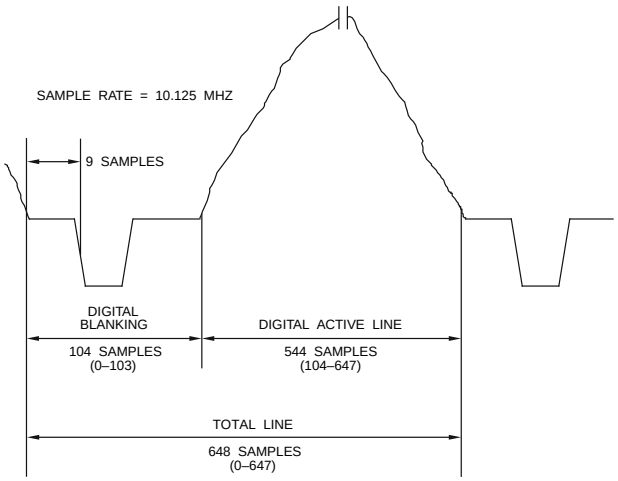
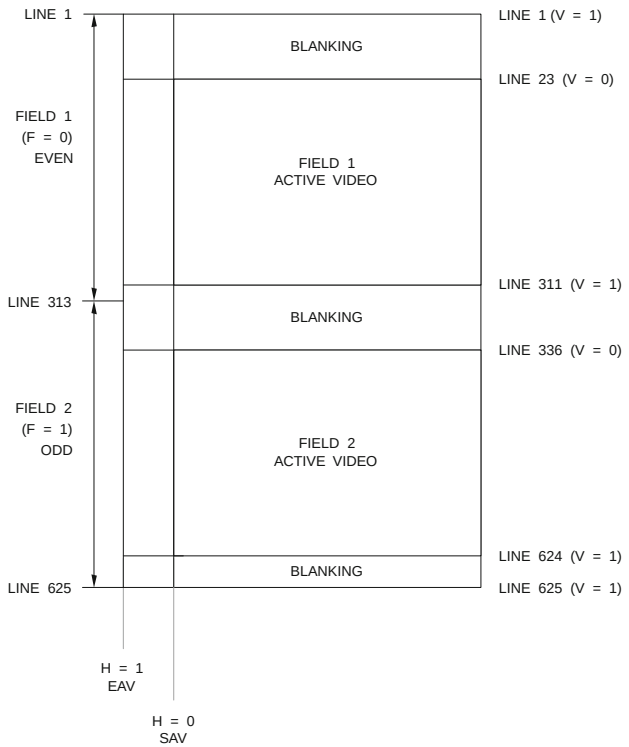


Figure 7.3-19 576i Analog–Digital Relationship (4:3 Aspect Ratio, 25 Hz Frame Rate, 10.125 MHz Sample Clock).

1024 × 576p	39.33 MHz	50 Hz
768 × 576p	29.50 MHz	50 Hz
704 × 576p	27.00 MHz	50 Hz
544 × 576p	20.25 MHz	50 Hz
480 × 576p	18.00 MHz	50 Hz

1024 × 576p is a 16:9 square pixel format, while 768 × 576p is a 4:3 square pixel format. The 704 × 576p format is done by using the 720 x 576p format, and blanking the first eight and last eight samples each active scan line. Example relationships between the analog and digital signals are shown in Figures 7.3-21 through 7.3-24.

The H (horizontal blanking), V (vertical blanking), and F (field) signals are defined in Figure 7.3-25. The H,



LINE NUMBER	F	V
1–22	0	1
23–310	0	0
311–312	0	1
313–335	1	1
336–623	1	0
624–625	1	1

Figure 7.3-20 576i Digital Vertical Timing (576 Active Lines). F and V change state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-16 through 7.3-19. IEC 61834-2, ITU-R BT.1618, and SMPTE 314M (DV formats) specify lines 23–310 and 335–622 for active video.

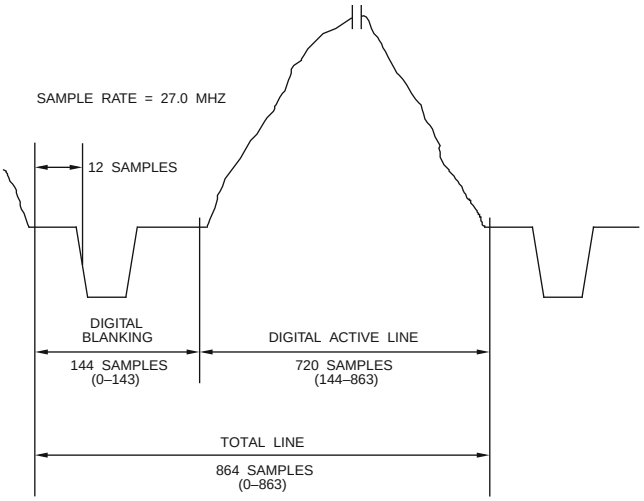


Figure 7.3-21 576p Analog–Digital Relationship (4:3 Aspect Ratio, 50 Hz Frame Rate, 27 MHz Sample Clock).

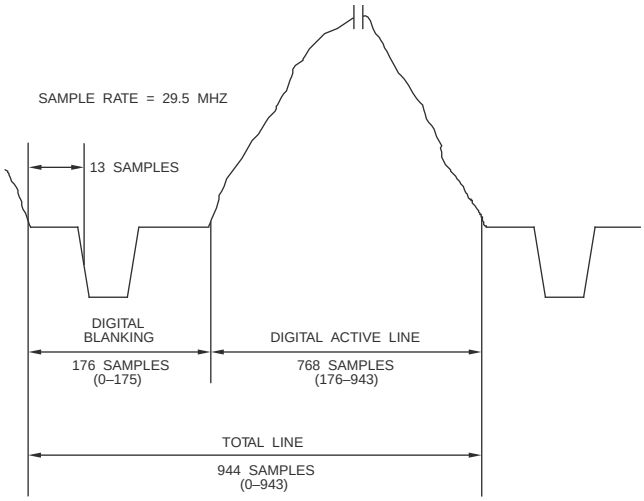


Figure 7.3-23 576p Analog–Digital Relationship (4:3 Aspect Ratio, 50 Hz Frame Rate, 29.5 MHz Sample Clock).

V, and F timing indicated is compatible with video compression standards rather than BT.656.

720p Systems

Progressive analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 720p (since there are typically 720 active scan lines per frame and they are progressive), the frame rate is usually 59.94 Hz (60/1.001) to simplify the generation of (M) NTSC video. The analog interface uses 750 lines per frame, with active video present on lines 26–745, as shown in Figure 7.3-26.

For the 59.94 Hz frame rate, each scan line time (H) is about 22.24 µs. Detailed horizontal timing is dependent on the specific video interface used.

Progressive digital component video

SMPTE 296M specifies the representation for 720p digital R′G′B′ or YCbCr signals. Active resolutions defined within SMPTE 296M, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

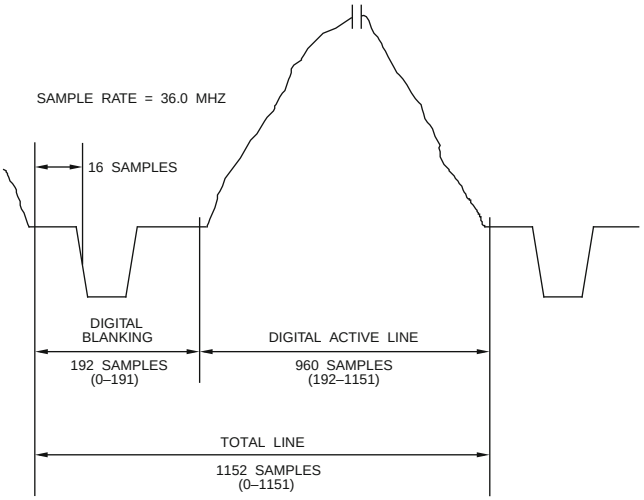


Figure 7.3-22 576p Analog–Digital Relationship (16:9 Aspect Ratio, 50 Hz Frame Rate, 36 MHz Sample Clock).

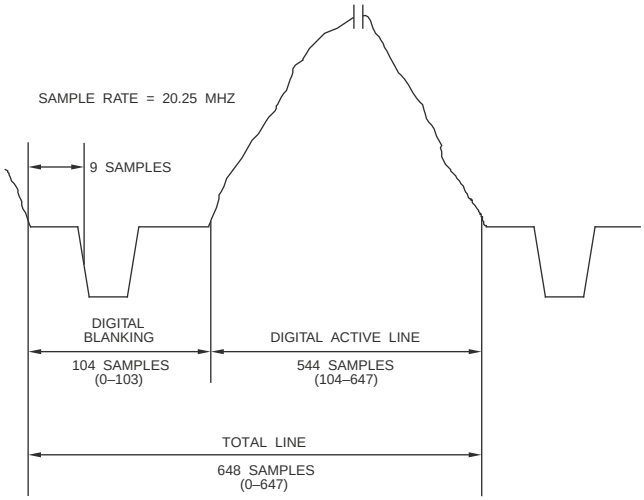
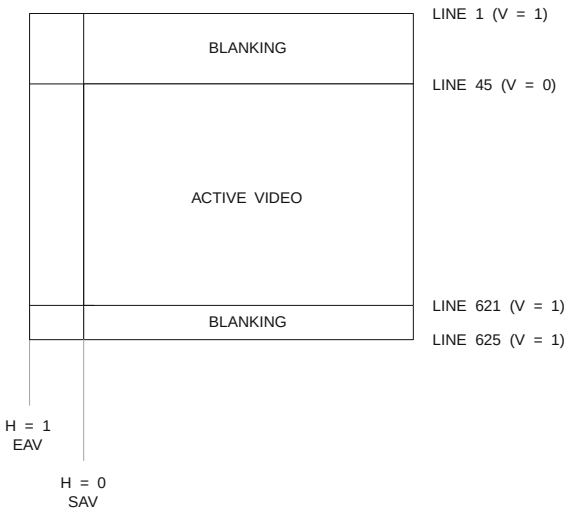


Figure 7.3-24 576p Analog–Digital Relationship (4:3 Aspect Ratio, 50 Hz Frame Rate, 20.25 MHz Sample Clock).



LINE NUMBER	F	V
1–44	0	1
45–620	0	0
621–625	0	1

Figure 7.3-25 576p Digital Vertical Timing (576 Active Lines). V changes state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-21 through 7.3-24.

1280 × 720p	74.176 MHz	23.976 Hz
1280 × 720p	74.250 MHz	24.000 Hz
1280 × 720p	74.250 MHz	25.000 Hz
1280 × 720p	74.176 MHz	29.970 Hz
1280 × 720p	74.250 MHz	30.000 Hz
1280 × 720p	74.250 MHz	50.000 Hz
1280 × 720p	74.176 MHz	59.940 Hz
1280 × 720p	74.250 MHz	60.000 Hz

Note that square pixels and a 16:9 aspect ratio are used. Example relationships between the analog and digital signals are shown in Figures 7.3-27 and 7.3-28, and Table 7.3-1. The H (horizontal blanking), V (vertical blanking), and F (field) signals are as defined in Figure 7.3-29.

1080i and 1080p Systems

Interlaced analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 1080i (since there are typically 1080 active scan lines per frame and they are interlaced), the frame rate is usually 25 or 29.97 Hz (30/1.001) to simplify the generation of (B, D, G, H, I) PAL or (M) NTSC video. The analog interface uses 1125 lines per frame, with active video present on lines 21–560 and 584–1123, as shown in Figure 7.3-30.

MPEG-2 and MPEG-4 systems use 1088 lines, rather than 1080, in order to have a multiple of 32 scan lines per frame. In this case, an additional 4 lines per field after the active video are used.

For the 25 Hz frame rate, each scan line time is about 35.56 µs. For the 29.97 Hz frame rate, each scan line time is about 29.66 µs. Detailed horizontal timing is dependent on the specific video interface used.

1152i Format

The 1152i active (1250 total) line format is not a broadcast transmission format. However, it is being used as an analog interconnection standard from HD set-top boxes and DVD players to 50 Hz CRT-based HDTVs. This enables 50 Hz HDTVs to use a fixed 31.25 kHz horizontal frequency, reducing their cost. Other HDTV display technologies, such as DLP, LCD, and plasma, are capable of handling the native timing of 720p50

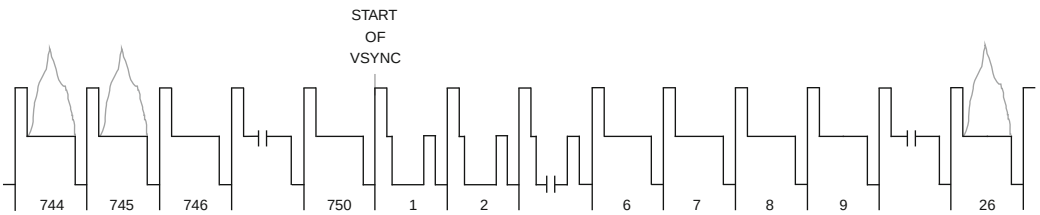


Figure 7.3-26 720p Vertical Interval Timing.

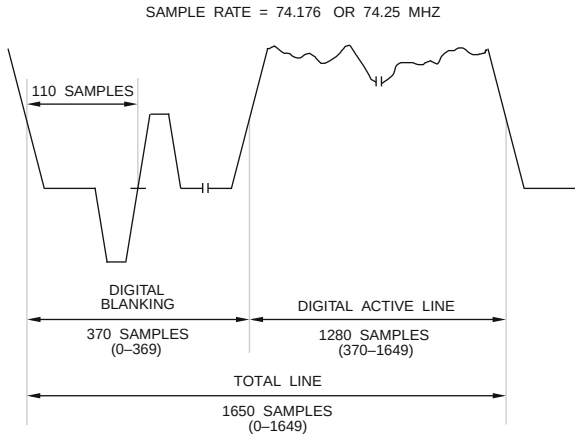


Figure 7.3-27 720p Analog–Digital Relationship (16:9 Aspect Ratio, 59.94 Hz Frame Rate, 74.176 MHz Sample Clock and 60 Hz Frame Rate, 74.25 MHz Sample Clock).

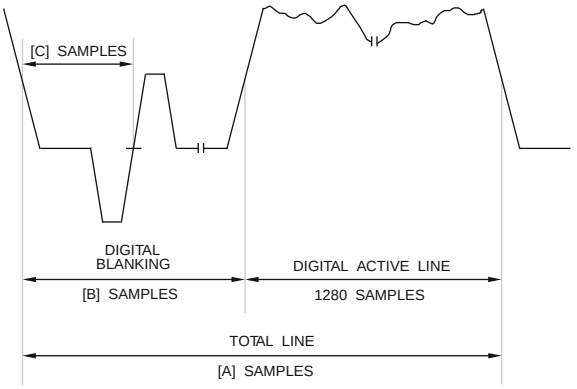


Figure 7.3-28 General 720p Analog–Digital Relationship.

Table 7.3-1 Various 720p analog–digital parameters for Figure 7.3-28.

Active horizontal samples	Frame rate (Hz)	1× Y Sample rate (MHz)	Total horizontal samples (A)	Horizontal blanking samples (B)	C samples
1280	24/1.001	74.25/1.001	4125	2845	2585
	24	74.25	4125	2845	2585
	25[1]	48	1536	256	21
	25[1]	49.5	1584	304	25
	25	74.25	3960	2680	2420
	30/1.001	74.25/1.001	3300	2020	1760
	30	74.25	3300	2020	1760
	50	74.25	1980	700	440
	60/1.001	74.25/1.001	1650	370	110
	60	74.25	1650	370	110

Note:
1. Useful for CRT-based 50 Hz HDTVs based on a 31.250 kHz horizontal frequency. Sync pulses are −300 mV bi-level, rather than ±300 mV tri-level. 720p content scaled vertically to 1152i active scan lines; 1250i total scan lines instead of 750p.

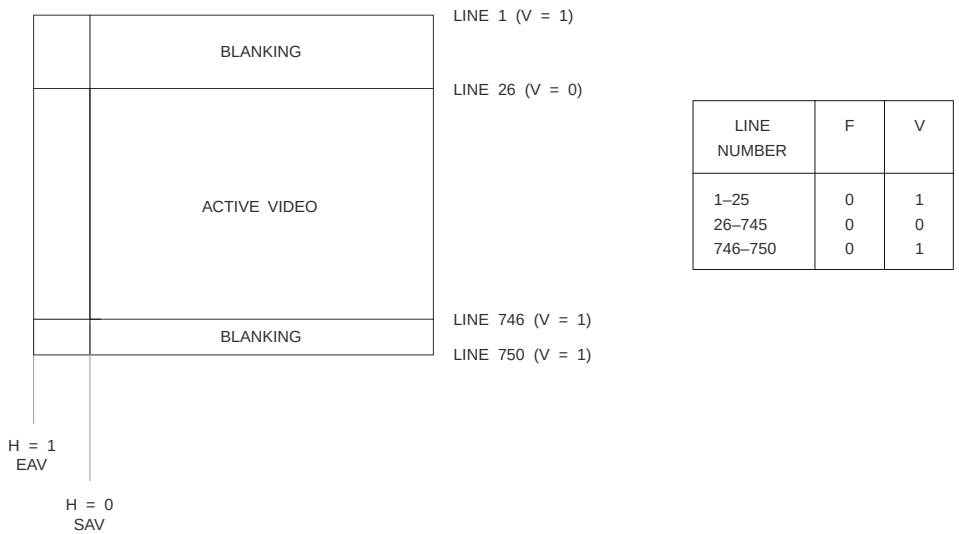


Figure 7.3-29 720p Digital Vertical Timing (720 Active Lines). V changes state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-27 and 7.3-28.

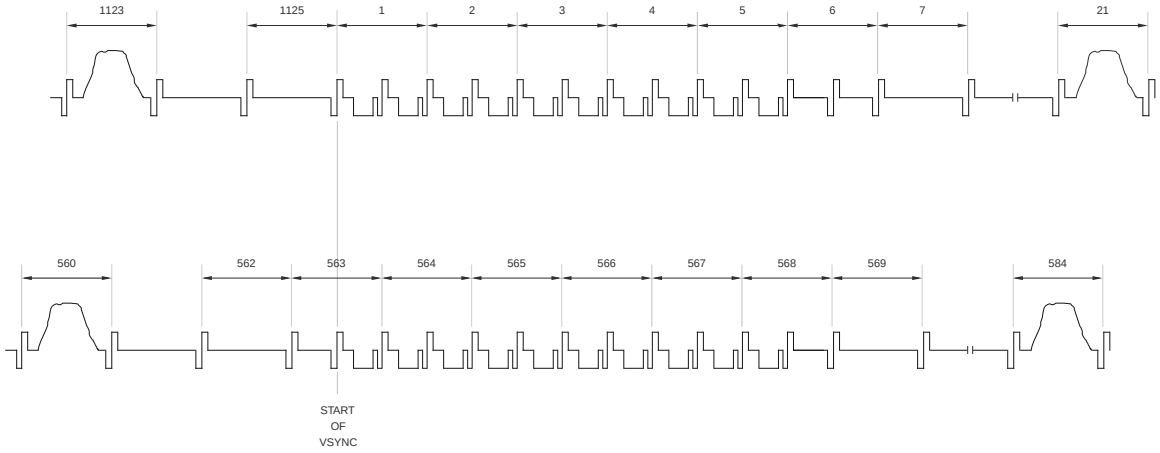


Figure 7.3-30 1080i Vertical Interval Timing.

(750p50 with VBI) and 1080i25 (1125i25 with VBI) analog signals.

The set-top box or DVD player converts 720p50 and 1080i25 content to the 1152i25 format. 1280 × 720p50 content is scaled to 1280 × 1152i25; 1920 × 1080i25 content is presented letterboxed in a 1920 × 1152i25 format. HDTVs will have a nominal vertical zoom mode for correcting the geometry of 1080i25, which can be recognized by the vertical synchronizing signal.

Progressive analog component video

Analog component signals are comprised of three signals, analog R′G′B′ or YPbPr. Referred to as 1080p (since there are typically 1080 active scan lines per frame and they are progressive), the frame rate is usually 50 or 59.94 Hz (60/1.001) to simplify the generation of (B, D, G, H, I) PAL or (M) NTSC video. The analog interface uses 1125 lines per frame, with active video present on lines 42–1121, as shown in Figure 7.3-31.

MPEG-2 and MPEG-4 systems use 1088 lines, rather than 1080, in order to have a multiple of 16 scan lines per frame. In this case, an additional 8 lines per frame after the active video are used.

For the 50 Hz frame rate, each scan line time is about 17.78 µs. For the 59.94 Hz frame rate, each scan line time is about 14.83 µs. Detailed horizontal timing is dependent on the specific video interface used.

Interlaced digital component video

ITU-R BT.709 and SMPTE 274M specify the digital component format for the 1080i digital R′G′B′ or YCbCr signal. Active resolutions defined within BT.709 and SMPTE 274M, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

1920 × 1080i	74.250 MHz	25.00 Hz
1920 × 1080i	74.176 MHz	29.97 Hz
1920 × 1080i	74.250 MHz	30.00 Hz

Note that square pixels and a 16:9 aspect ratio are used. Other common active resolutions, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

1280 × 1080i	49.500 MHz	25.00 Hz
1280 × 1080i	49.451 MHz	29.97 Hz
1280 × 1080i	49.500 MHz	30.00 Hz
1440 × 1080i	55.688 MHz	25.00 Hz
1440 × 1080i	55.632 MHz	29.97 Hz
1440 × 1080i	55.688 MHz	30.00 Hz

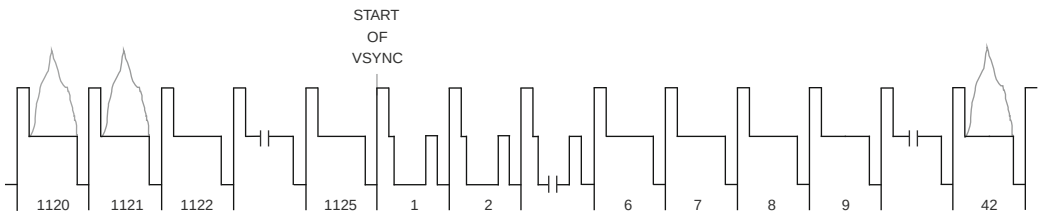


Figure 7.3-31 1080p Vertical Interval Timing.

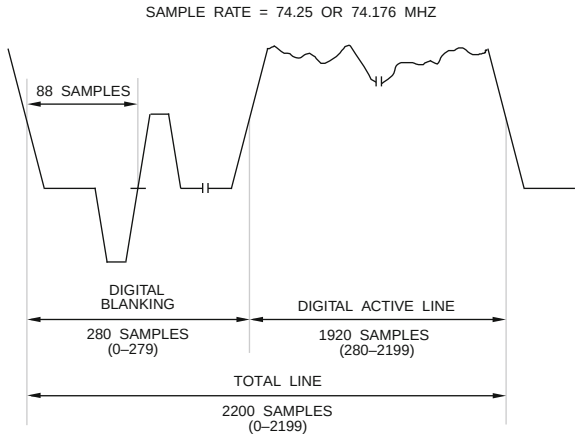


Figure 7.3-32 1080i Analog–Digital Relationship (16:9 Aspect Ratio, 29.97 Hz Frame Rate, 74.176 MHz Sample Clock and 30 Hz Frame Rate, 74.25 MHz Sample Clock).

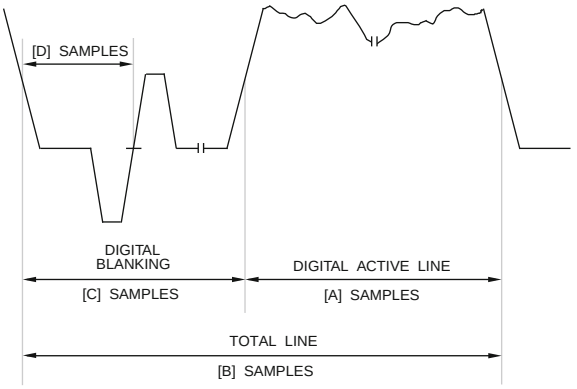


Figure 7.3-33 General 1080i Analog–Digital Relationship.

Example relationships between the analog and digital signals are shown in Figures 7.3-32 and 7.3-33, and Table 7.3-2. The H (horizontal blanking) and V (vertical blanking) signals are as defined in Figure 7.3-34.

Progressive digital component video

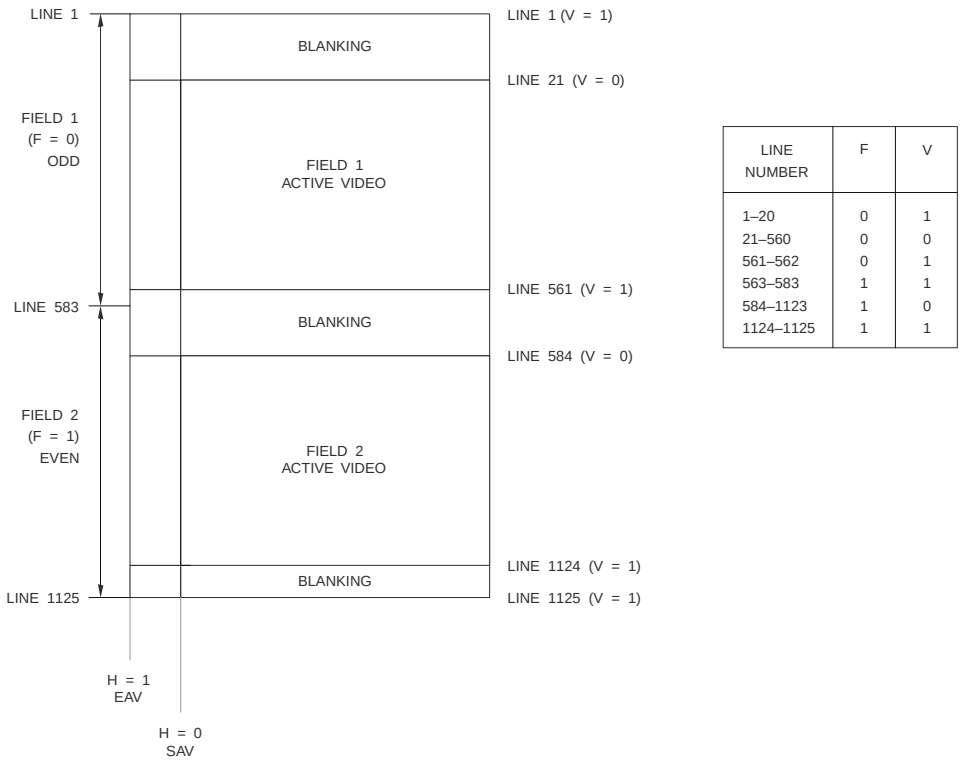
ITU-R BT.709 and SMPTE 274M specify the digital component format for the 1080p digital R′G′B′ or YCbCr signal. Active resolutions defined within BT.709 and SMPTE 274M, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

1920 × 1080p	74.176 MHz	23.976 Hz
1920 × 1080p	74.250 MHz	24.000 Hz
1920 × 1080p	74.250 MHz	25.000 Hz
1920 × 1080p	74.176 MHz	29.970 Hz
1920 × 1080p	74.250 MHz	30.000 Hz
1920 × 1080p	148.50 MHz	50.000 Hz
1920 × 1080p	148.35 MHz	59.940 Hz
1920 × 1080p	148.50 MHz	60.000 Hz

Table 7.3-2 Various 1080i analog–digital parameters for Figure 7.3-33.

Active horizontal samples (A)	Frame rate (Hz)	1× Y sample rate (MHz)	Total horizontal samples (B)	Horizontal blanking samples (C)	D samples
1920	25[1] 25[1] 25 30/1.001 30	72 74.25 74.25 74.25/1.001 74.25	2304 2376 2640 2200 2200	384 456 720 280 280	32 38 528 88 88
1440	25[1] 25 30/1.001 30	54 55.6875 55.6875/1.001 55.6875	1728 1980 1650 1650	288 540 210 210	24 396 66 66
1280	25[1] 25 30/1.001 30	48 49.5 49.5/1.001 49.5	1536 1760 1466.7 1466.7	256 480 186.7 186.7	21 352 58.7 58.7

Note:
1. Useful for CRT-based 50 Hz HDTVs based on a 31.250 kHz horizontal frequency. Sync pulses are −300 mV bi-level, rather than ±300 mV tri-level. 1080i content letterboxed in 1152i active scan lines; 1250i total scan lines instead of 1125i.



LINE NUMBER	F	V
1–20	0	1
21–560	0	0
561–562	0	1
563–583	1	1
584–1123	1	0
1124–1125	1	1

Figure 7.3-34 1080i Digital Vertical Timing (1080 Active Lines). F and V change state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-32 and 7.3-33.

Note that square pixels and a 16:9 aspect ratio are used. Other common active resolutions, their 1× Y and R′G′B′ sample rates (F_s), and frame rates, are:

1280 × 1080p	49.451 MHz	23.976 Hz
1280 × 1080p	49.500 MHz	24.000 Hz
1280 × 1080p	49.500 MHz	25.000 Hz
1280 × 1080p	49.451 MHz	29.970 Hz
1280 × 1080p	49.500 MHz	30.000 Hz
1280 × 1080p	99.000 MHz	50.000 Hz
1280 × 1080p	98.901 MHz	59.940 Hz
1280 × 1080p	99.000 MHz	60.000 Hz
1440 × 1080p	55.632 MHz	23.976 Hz
1440 × 1080p	55.688 MHz	24.000 Hz
1440 × 1080p	55.688 MHz	25.000 Hz
1440 × 1080p	55.632 MHz	29.970 Hz
1440 × 1080p	55.688 MHz	30.000 Hz
1440 × 1080p	111.38 MHz	50.000 Hz
1440 × 1080p	111.26 MHz	59.940 Hz
1440 × 1080p	111.38 MHz	60.000 Hz

Example relationships between the analog and digital signals are shown in Figures 7.3-35 and 7.3-36, and Table 7.3-3. The H (horizontal blanking), V (vertical blanking), and F (field) signals are as defined in Figure 7.3-37.

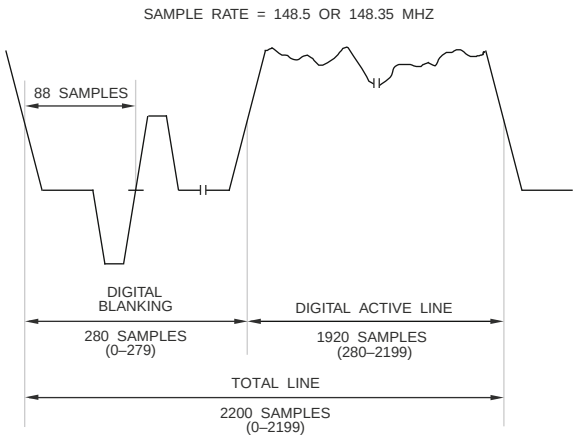


Figure 7.3-35 1080p Analog–Digital Relationship (16:9 Aspect Ratio, 59.94 Hz Frame Rate, 148.35 MHz Sample Clock and 60 Hz Frame Rate, 148.5 MHz Sample Clock).

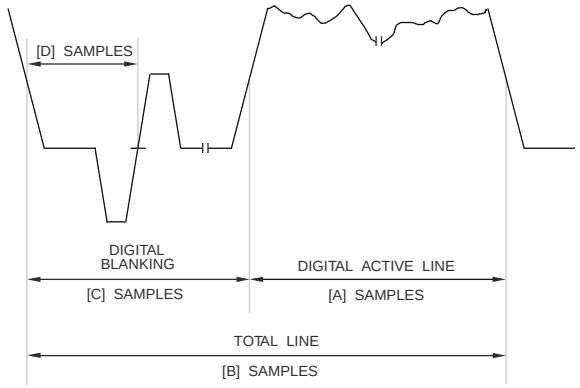


Figure 7.3-36 General 1080p Analog–Digital Relationship.

640 × 400	VGA
640 × 480	VGA
854 × 480	WVVGA
800 × 600	SVGA
1024 × 768	XGA
1280 × 768	WXGA
1366 × 768	WXGA
1024 × 1024	XGA
1280 × 1024	SXGA
1600 × 1024	WSXGA
1600 × 1200	UXGA
1920 × 1200	WUXGA

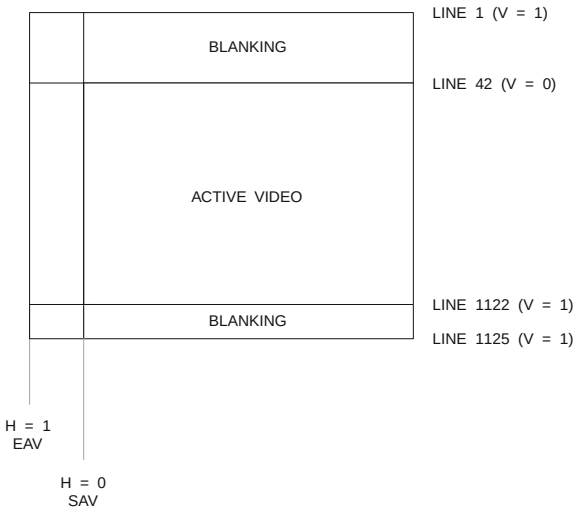
Other video systems

Some consumer displays, such as those based on LCD and plasma technologies, have adapted other resolutions as their native resolution. Common active resolutions and their names are:

These resolutions, and their timings, are defined for computer monitors by the Video Electronics Standards Association (VESA). Displays based on one of these native resolutions are usually capable of accepting many input resolutions, scaling the source to match the display resolution.

Table 7.3-3 Various 1080p analog–digital parameters for Figure 7.3-36.

Active horizontal samples (A)	Frame rate (Hz)	1× Y sample rate (MHz)	Total horizontal samples (B)	Horizontal blanking samples (C)	D samples
1920	24/1.001	74.25/1.001	2750	830	638
	24	74.25	2750	830	638
	25	74.25	2640	720	528
	30/1.001	74.25/1.001	2200	280	88
	30	74.25	2200	280	88
	50	148.5	2640	720	528
	60/1.001	148.5/1.001	2200	280	88
	60	148.5	2200	280	88
1440	24/1.001	55.6875/1.001	2062.5	622.5	478.5
	24	55.6875	2062.5	622.5	478.5
	25	55.6875	1980	540	396
	30/1.001	55.6875/1.001	1650	210	66
	30	55.6875	1650	210	66
	50	111.375	1980	540	396
	60/1.001	111.375/1.001	1650	210	66
	60	111.375	1650	210	66
1280	24/1.001	49.5/1.001	1833.3	553.3	425.3
	24	49.5	1833.3	553.3	425.3
	25	49.5	1760	480	352
	30/1.001	49.5/1.001	1466.7	186.7	58.7
	30	49.5	1466.7	186.7	58.7
	50	99	1760	480	352
	60/1.001	99/1.001	1466.7	186.7	58.7
	60	99	1466.7	186.7	58.7



LINE NUMBER	F	V
1–41	0	1
42–1121	0	0
1122–1125	0	1

Figure 7.3-37 1080pDigital Vertical Timing (1080 Active Lines). V changes state at the EAV sequence at the beginning of the digital line. Note that the digital line number changes state prior to the start of horizontal sync, as shown in Figures 7.3-35 and 7.3-36.

References

1. CEA-861D, *A DTV Profile for Uncompressed High Speed Digital Interfaces*, July 2006.
2. EIA-770.1, *Analog 525-Line Component Video Interface—Three Channels*, November 2001.
3. EIA-770.2, *Standard-Definition TV Analog Component Video Interface*, November 2001.
4. EIA-770.3, *High-Definition TV Analog Component Video Interface*, November 2001.
5. ITU-R BT.601–5, 1995, *Studio Encoding Parameters of Digital Television for Standard 4:3 and Widescreen 16:9 Aspec Ratios.*
6. ITU-R BT.709–5, 2002, *Parameter Values for the HDTV Standards for Production and International Programme Exchange.*
7. ITU-R BT.1358, 1998, *Studio Parameters of 625 and 525 Line Progressive Scan Television Systems.*
8. SMPTE 267M–1995, *Television—Bit-Parallel Digital Interface*—Component Video Signal 4:2:216 × 9 Aspect Ratio.
9. SMPTE 274M–2005, *Television—1920 × 1080 Image Sample Structure, Digital Representation and Digital Timing Reference Sequences for Multiple Picture Rates.*
10. SMPTE 293M–2003, *Television—720 × 483 Active Line at 59.94 Hz Progressive Scan Production—Digital Representation.*
11. SMPTE 296M–2001, *Television—1280 × 720 Progressive Image Sample Structure, Analog and Digital Representation and Analog Interface.*

This page intentionally left blank

Chapter 7.4

7.4

Video compression

Alan Bovik

7.4.1 Introduction

The subject of video coding is of fundamental importance to many areas in engineering and the natural and perceptual sciences. Video engineering is quickly becoming a largely digital discipline although analog TV transmission is still by far the mainstay around the world. Digital transmission of television signals via satellites is commonplace, and widespread HDTV terrestrial transmission began in 2000 and is targeted to become the default transmission standard in the United States by 2006. Video compression is an absolute requirement for the growth and success for the low bandwidth transmission and storage of digital video signals. Video encoding is used wherever digital video communications, storage, processing, acquisition, and reproduction occur. The transmission of high quality multimedia information over high speed computer networks is a central problem in the design of *quality of services* (QoS) for digital transmission providers. The *Motion Pictures Expert Group* (MPEG) finalized four well known encoding standards, MPEG-1, MPEG-2, MPEG-4, and MPEG-7. MPEG-1 and MPEG-2 define methods for the transmission of digital video information for multimedia and television formats. The MPEG-4 standard, which was adopted in 1999, specifically addresses the transmission of very low bit rate video by introducing the notion of media objects which are made up of audio, visual, or audiovisual content. It is targeted to satisfy the needs of audiovisual content authors, service providers, and end users. The MPEG-7 standard, which was adopted in 2001, defines audiovisual content storage and retrieval services. An aspect central to each of the MPEG standards are the video encoding and decoding algorithms that make digital video applications practical.

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2005 Elsevier Inc. All rights reserved

Video compression not only reduces the storage requirements or transmission bandwidth of digital video applications, but also affects many system performance tradeoffs. The design and selection of a video encoder therefore is not based only on its ability to compress information. Issues such as bit rate versus distortion criteria, algorithm complexity, transmission channel characteristics, algorithm symmetry versus asymmetry, video source statistics, fixed versus variable rate coding, and standards compatibility should be considered in order to make good encoder design decisions.

The growth of digital video applications and technology in the last few years has been explosive, and video compression is playing a central role in this success. Yet, the video coding discipline is relatively young and certainly will evolve and change significantly over the next few years. For instance, perceptual based video encoding research is still relatively young and promises to be able to significantly influence the future course of the discipline. Research in video coding has great vitality and the body of work is significant. It is apparent that this relevant and important topic will have an immense affect on the future of digital video technologies.

7.4.2 Introduction to video compression

Video or visual communications require significant amounts of information transmission. Video compression as considered here, involves the bit rate reduction of the digital video signal carrying visual information. Traditional video based compression like other information compression techniques, focuses on eliminating redundancy and unimportant elements of the source. The

degree to which the encoder reduces the bit rate is called its *coding efficiency*, or equivalently its inverse is termed the *compression ratio*, i.e.,

$$\begin{aligned}\text{coding efficiency} &= (\text{compression ratio})^{-1} \\ &= \text{encoded bit rate}/\text{decoded bit rate}.\end{aligned} \tag{7.4.1}$$

Compression can be a lossless or lossy operation. Due to the immense volume of video information, lossy operations are a key element used in video compression algorithms. The loss of information or distortion measure is usually evaluated using the *mean square error* (MSE), *mean absolute error* (MAE) criteria, or *peak signal-to-reconstruction noise* (PSNR),

$$\begin{aligned} MSE &= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}[I(i,j)-\hat{I}(i,j)]^2 \\ MAE &= \frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left|I(i,j)-\hat{I}(i,j)\right| \\ PSNR &= 20\log_{10}\left(\frac{2^n}{MSE^{1/2}}\right), \end{aligned} \tag{7.4.2}$$

for image I and its reconstructed image $\hat{I}$, pixel indices $1 \le i \le M$ and $1 \le j \le N$, image size $N \times M$ pixels, and n bits per pixel. The MSE, MAE, and PSNR as described here are global measures and do not necessarily give a good indication of the reconstructed image quality. In the final analysis, the human observer determines the quality of the reconstructed image and video quality. The concept of distortion versus coding efficiency is one of the most fundamental tradeoffs in the technical evaluation of video encoders.

Video signals contain information in three dimensions. These dimensions are modeled as *spatial* and *temporal* domains for video encoding purposes. Digital video compression methods generally seek to minimize information redundancy independently in each domain. The major international video compression standards (MPEG-1, MPEG-2, MPEG-4, H.261, H.262, and H.263) use this approach. Figure 7.4-1 schematically depicts a generalized video compression system that implements the spatial and temporal encoding of a digital image sequence. Each image in the sequence I_k is defined as in Eq. (7.4.2). The spatial encoder operates on image *blocks*, typically on the order of 8×8 pixels each. The temporal encoder generally operates on 16×16 pixel image blocks. The system is designed for two modes of operation, the *intraframe* mode and the *interframe* mode.

The single layer feedback structure of this generalized model is representative of the encoders that are recommended by the *International Standards Organization* (ISO) and *International Telecommunications Union* (ITU) video coding standards MPEG-1, MPEG-2/H.262, MPEG-4, and H.261 [1, 2, 3, 4]. The feedback loop is used in the Interframe mode of operation and generates a *prediction error* between the image blocks of the current frame and the current prediction frame. The prediction is generated by the *motion compensator*. The *motion estimation* unit creates *motion vectors* for each 16×16 block. The motion vectors and previously reconstructed frame are fed to the motion compensator to create the prediction.

The intraframe mode spatially encodes an entire current frame on a periodic basis, such as every 15 frames, to ensure that systematic errors do not continuously propagate. Intraframe mode will also be used to spatially encode a block whenever the Interframe encoding mode cannot meet its performance threshold. The intraframe

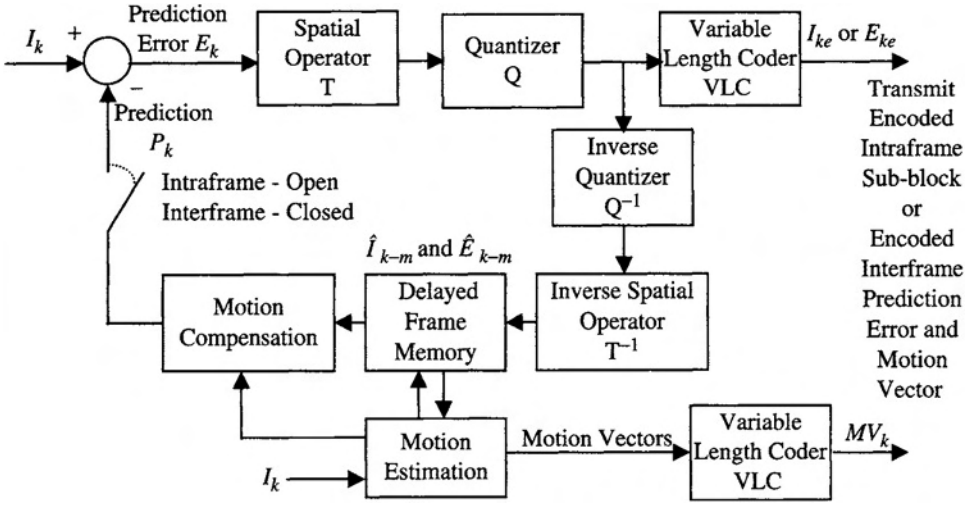


Figure 7.4-1 Generalized video compression system.

versus interframe mode selection algorithm is not included in this diagram. It is responsible for controlling the selection of the encoding functions, data flows, and output data streams for each mode.

The intraframe encoding mode does not receive any input from the feedback loop. I_k is spatially encoded, and subsequently encoded by the VLC forming I_{ke}, which is transmitted to the decoder. The receiver decodes I_{ke} producing the reconstructed image sub-block $\widehat{I}_k$. During the interframe coding mode, the current frame prediction P_k is subtracted from the current frame input I_k to form the current prediction error E_k. The prediction error is then spatially and VLC encoded to form E_{ke} and is transmitted along with the VLC encoded motion vectors MV_k. The decoder can reconstruct the current frame $\widehat{I}_k$ using the previously reconstructed frame $\widehat{I}_{k-1}$ (stored in the decoder), the current frame motion vectors, and the prediction error. The motions vectors MV_k operate on $\widehat{I}_{k-1}$ to generate the current prediction frame P_k. The encoded prediction error E_{ke} is decoded to produce the reconstructed prediction error $\widehat{E}_k$. The prediction error is added to the prediction to form the current frame $\widehat{I}_k$. The functional elements of the generalized model are described here in detail.

- **Spatial operator** — This element is generally a unitary two dimensional linear transform, but in principle can be any unitary operator that can distribute most of the signal energy into a small number of coefficients, i.e., decorrelate the signal data. Spatial transformations are successively applied to small image blocks in order to take advantage of the high degree of data correlation in adjacent image pixels. The most widely used spatial operator for image and video coding is the *discrete cosine transform* (DCT). It is applied to 8×8 pixel image blocks, and is well suited for image transformations because it uses real computations with fast implementations, provides excellent decorrelation of signal components, and avoids generation of spurious components between the edges of adjacent image blocks.

- **Quantizer** — The spatial or transform operator is applied to the input in order to arrange the signal into a more suitable format for subsequent lossy and lossless coding operations. The quantizer operates on the transform generated coefficients. This is a lossy operation that can result in a significant reduction in the bit rate. The quantization method used in this kind of video encoder is usually scalar and non-uniform. The scalar quantizer simplifies the complexity of the operation as compared to *vector quantization* (VQ). The non-uniform quantization interval is sized according to the distribution of the transform coefficients in order to minimize the bit rate and the distortion created by the quantization process. Alternatively, the quantization interval size can be adjusted based on the performance of the *human visual system* (HVS). The *Joint Pictures Expert Group* (JPEG) standard includes two (luminance and color difference) HVS sensitivity weighted quantization matrices in its "Examples and Guidelines" annex.

- **Variable length coding** — The lossless variable length coding (VLC) is used to effectively exploit the "symbolic" redundancies contained in each block of quantized transform coefficients. This step is termed "entropy coding" to designate that the encoder is designed to minimize the source entropy. The VLC is applied to a serial bit stream that is generated by scanning the transform coefficient block. The scanning pattern should be chosen with the objective of maximizing the performance of the VLC. The MPEG encoder for instance, describes a zigzag scanning pattern that is intended to maximize transform zero coefficient run lengths. The H.261 VLC is designed to encode these run lengths using a variable length *Huffman* code.

The feedback loop sequentially reconstructs the encoded spatial and prediction error frames and stores the results in order to create a prediction for the current image sub-block. The elements required to do this are the inverse quantizer, inverse spatial operator, delayed frame memory, motion estimator, and motion compensator.

- **Inverse operators** — The inverse operators Q^{-1} and T^{-1} are applied to the encoded current frame I_{ke} or the current prediction error E_{ke} in order to reconstruct and store the encoded frame for the motion estimator and motion compensator to generate the next prediction frame.
- **Delayed frame memory** — Both current and previous frames must be available to the motion estimator and motion compensator to generate a prediction frame. The number of previous frames stored in memory can vary based upon the requirements of the encoding algorithm. MPEG-1 defines a B-frame which is a bidirectional encoding that requires that motion prediction be performed in both the forward and backward directions. This necessitates storage of multiple frames in memory.
- **Motion estimation** — The temporal encoding aspect of this system relies on the assumption that rigid body motion is responsible for the differences between two or more successive frames. The objective of the motion estimator is to estimate the rigid body motion between two frames. The motion estimator operates on all current frame 16×16 image blocks and generates the pixel displacement or *motion vector* for each block. The technique used to generate motion vectors is called *block-matching motion*

estimation. The method uses the current frame I_k and the previous reconstructed frame $\hat{I}_{k-1}$ as input. Each block in the previous frame is assumed to have a displacement that can be found by searching for it in the current frame. The search is usually constrained to be within a reasonable neighborhood so as to minimize the complexity of the operation. Search matching is usually based on a minimum MSE or MAE criteria. When a match is found, the pixel displacement is used to encode the particular block. If a search does not meet a minimum MSE or MAE threshold criteria, the motion compensator will indicate that the current block is to be spatially encoded using Intraframe mode.

- **Motion compensation** — The motion compensator makes use of the current frame motion estimates MV_k and the previously reconstructed frame $\hat{I}_{k-1}$ to generate the current frame prediction P_k. The current frame prediction is constructed by placing the previous frame blocks into the current frame according to the motion estimate pixel displacement. The motion compensator uses the threshold criteria to decide which blocks will be encoded as prediction error blocks using motion vectors and which blocks will only be spatially encoded.

The generalized model does not address some video compression system details such as the bit-stream syntax (which supports different application requirements), or the specifics of the encoding algorithms. These issues are dependent upon the video compression system design.

Alternative video encoding models have also been the focus of current research. Three-dimensional (3D) video information can be compressed directly using VQ or 3D *wavelet* encoding models. VQ encodes a 3D block of pixels as a codebook index that denotes its "closest or nearest neighbor" in the minimum squared or absolute error sense. However, the VQ codebook size grows on the order as the number of possible inputs. Searching the codebook space for the nearest neighbor is generally very computationally complex, but structured search techniques can provide good bit rates, quality, and computational performance. *Tree-structured* VQ (TSVQ) [14] reduces the search complexity from codebook size N to log N, with a corresponding loss in average distortion. The simplicity of the VQ decoder (it only requires a simple table lookup for the transmitted codebook index), and its bit rate-distortion performance make it an attractive alternative for specialized applications. The complexity of the codebook search generally limits the use of VQ in real-time applications. VQ quantizers have also been proposed for Interframe, variable bit rate, and subband video compression methods [5].

Three dimensional wavelet encoding is a topic of recent interest. This video encoding method is based on the *discrete wavelet transform* methods. The wavelet transform decomposes a bandwidth limited signal into a *multiresolution* representation. The multiresolution decomposition makes the wavelet transform an excellent signal analysis tool because signal characteristics can be viewed in a variety of time-frequency or space-frequency scales. The wavelet transform is implemented in practice via the use of multiresolution or *subband* filterbanks [6]. The wavelet filterbank is well suited for video encoding because of its ability to adapt to the multiresolution characteristics of video signals. Wavelet transform encodings are naturally hierarchic in their time-frequency representation and easily adaptable for *progressive transmission* [7]. They have also been shown to possess excellent bit rate-distortion characteristics.

Direct three dimensional video compression systems suffer from a major drawback for real-time encoding and transmission. In order to encode a sequence of images in one operation, the sequence must be buffered. This introduces a buffering and computational delay that can be very noticeable in the case of real-time video communications.

Video compression techniques treating visual information in accordance with *human visual system* models have recently been introduced. These methods are termed "second-generation or object-based," and attempt to achieve very large compression ratios by imitating the operations of the HVS. The HVS model has also been incorporated into more traditional video compression techniques by reflecting visual perception into various aspects of the coding algorithm. HVS weightings have been designed for the DCT AC coefficients quantizer used in the MPEG encoder.

Digital video compression is currently enjoying tremendous growth in part due to the great advances in VLSI, ASIC, and microcomputer technology in the last decade. The realtime nature of video communications necessitates the use of general purpose and specialized high performance hardware devices. In the near future, advances in design and manufacturing technologies will create hardware devices that will allow greater adaptability, interactivity, and interoperability of video applications. For instance, MPEG-7 has defined format free operations for the storage and retrieval of audio-visual information that is being used by digital cable TV vendors for "on demand" delivery of digital video content.

7.4.3 Video compression application requirements

A wide variety of digital video applications currently exist. They range from simple low resolution and bandwidth applications (multimedia, PicturePhone) to very high resolution and bandwidth (HDTV) demands. This

section will present requirements of current and future digital video applications and the demands they place on the video compression system.

To demonstrate the importance of video compression, the transmission of digital video television signals is presented. The bandwidth required by a digital television signal is approximately one-half the number of picture elements (pixels) displayed per second. The analog video monitor pixel size in the vertical dimension is the distance between scanning lines, and the horizontal dimension is the distance the scanning spot moves during 1/2 cycle of the highest video signal transmission frequency. The video signal bandwidth is given by Eq. (7.4.3),

$$\begin{aligned} B_w &= (\text{cycles/frame})(F_R) \\ &= (\text{cycles/line})(N_L)(F_R) \\ &= \frac{(0.5)(\text{aspect ratio})(F_R)(N_L)(R_H)}{0.84} \\ &= (0.8)(F_R)(N_L)(R_H) \end{aligned} \tag{7.4.3}$$

where

B_w = video signal system bandwidth
F_R = number of frames transmitted per second (fps)
N_L = number of scanning lines per frame
R_H = horizontal resolution (lines), proportional to pixel resolution

The NTSC picture aspect ratio is 4/3, the constant 0.5 is the ratio of the number of cycles to the number of lines, and the factor 0.84 is the fraction of the horizontal scanning interval that is devoted to signal transmission.

The *National Television Systems Committee* (NTSC) transmission standard used for television broadcasts in the United States has the following parameter values,

$$\begin{aligned} F_R &= 29.97 \text{ fps} \\ N_L &= 525 \text{ lines} \\ R_H &= 340 \text{ lines.} \end{aligned}$$

This yields an analog video system bandwidth B_w of 4.2 MHz for the NTSC broadcast system. In order to transmit a color digital video signal, the digital pixel format must be defined. The digital color pixel is made of three components: One luminance (Y) component occupying 8 bits, and two color difference components (U and V) each requiring 8 bits. The NTSC picture frame has 720 × 480 × 2 total luminance and color pixels. In order to transmit this information for an NTSC broadcast system at 29.97 frames per second, the following bandwidth is required,

$$\begin{aligned} \text{Digital BW} \simeq 1/2 \text{ bitrate} &= 1/2(29.97 \text{ fps}) \\ &\times (24 \text{ bits/pixel}) \times (720 \times 480 \\ &\times 2 \text{ pixels/frame}) = 249 \text{ MHz.} \end{aligned}$$

This represents an increase of approximately 59 times the required NTSC system bandwidth, and about 41 times the full transmission channel bandwidth (6 MHz) for current NTSC signals. HDTV picture resolution requires up to three times more raw bandwidth than this example! (Two transmission channels totaling 12 MHz are allocated for terrestrial HDTV transmissions.) It is clear from this example that terrestrial television broadcast systems have to use digital transmission and digital video compression to achieve the overall bit rate reduction and image quality required for HDTV signals.

The example not only points out the significant system bandwidth requirements for digital video information, but also indirectly brings up the issue of digital video quality requirements. The tradeoff between bit rate and quality or distortion is a fundamental issue facing the design of video compression systems. To this end, it is important to fully characterize an application's video communications requirements before designing or selecting an appropriate video compression system. Factors that should be considered in the design and selection of a video compression system include the following items:

- **Video characteristics** — Video parameters such as the dynamic range, source statistics, pixel resolution, and noise content can affect the performance of the compression system.
- **Transmission requirements** — Transmission bit rate requirements determine the power of the compression system. Very high transmission bandwidth, storage capacity, or quality requirements may necessitate lossless compression. Conversely, extremely low bit rate requirements may dictate compression systems that trade-off image quality for a large compression ratio. *Progressive transmission* is a key issue for selection of the compression system. It is generally used when the transmission bandwidth exceeds the compressed video bandwidth. Progressive coding refers to a multi-resolution, hierarchic, or subband encoding of the video information. It allows for transmission and reconstruction of each resolution independently from low to high resolution. Channel errors affect system performance and the quality of the reconstructed video. Channel errors can affect the bit stream randomly or in burst fashion. The channel error characteristics can have different effects on different encoders, and can range from local to global anomalies. In general, transmission error correction codes (ECC) are used to mitigate the effect of channel errors, but awareness and knowledge of this issue is important.
- **Compression system characteristics and performance** — The nature of video applications makes many demands on the video compression system.

Interactive video applications such as videoconferencing demand that the video compression systems have symmetric capabilities. That is, each participant in the interactive video session must have the same video encoding and decoding capabilities, and that the system performance requirements must be met by both the encoder and decoder. On the other hand, television broadcast video has significantly greater performance requirements at the transmitter because it has the responsibility of providing real-time high quality compressed video that meets the transmission channel capacity. Digital video system implementation requirements can vary significantly. Desktop televideo conferencing can be implemented using software encoding and decoding or may require specialized hardware and transmission capabilities to provide high quality performance. The characteristics of the application will dictate the suitability of the video compression algorithm for particular system implementations. The importance of the encoder and system implementation decision cannot be overstated, system architectures and performance capabilities are changing at a rapid pace and the choice of the best solution requires careful analysis of the all possible system and encoder alternatives.

- **Rate-distortion requirements** — The rate-distortion requirement is a basic consideration in the selection of the video encoder. The video encoder must be able to provide the bit rate(s) and video fidelity (or range of video fidelity) required by the application. Otherwise, any aspect of the system may not meet specifications. For example, if the bit rate specification is exceeded in order to support a lower MSE, a larger than expected transmission error rate may cause a catastrophic system failure.
- **Standards requirements** — Video encoder compatibility with existing and future standards is an important consideration if the digital video system is required to inter-operate with existing and/or future systems. A good example is that of a desktop videoconferencing application supporting a number of legacy video compression standards. This requires support of the older video encoding standards on new equipment designed for a newer incompatible standard. Videoconferencing equipment not supporting the old standards would not work in such environments.

These factors are displayed in Table 7.4-1 to demonstrate video compression system requirements for some common video communications applications. The video compression system designer as a minimum should consider these factors in making a determination about the choice of video encoding algorithms and technology to implement.

7.4.4 Digital video signals and formats

Video compression techniques make use of signal models in order to be able to utilize the body of digital signal analysis/processing theory and techniques that have been developed over the past fifty or so years. The design of a video compression system as represented by the generalized model introduced in Section 7.4.2, requires knowledge of the signal characteristics, and the digital processes that are used to create the digital video signal. It is also highly desirable to understand video display systems, and the behavior of the HVS.

7.4.4.1 Sampling of analog video signals

Digital video information is generated by sampling the *intensity* of the original continuous analog video signal $I(x, y, t)$ in three dimensions. The spatial component of the video signal is sampled in the horizontal and vertical dimensions (x, y), and the temporal component is sampled in the time dimension (t). This generates a series of digital images or image sequence $I(i, j, k)$. Video signals that contain color information are usually decomposed into three parameters (YC_rC_b, YUV, RGB, etc.) whose intensities are likewise sampled in three dimensions. The sampling process inherently quantizes the video signal due to the digital word precision used to represent the intensity values. Therefore the original analog signal can never be reproduced exactly, but for all intents and purposes, a high-quality digital video representation can be reproduced with arbitrary closeness to the original analog video signal.

An important result of sampling theory is the *Nyquist sampling theorem*. This theorem defines the conditions under which sampled analog signals can be "perfectly" reconstructed. If these conditions are not met, the resulting digital signal will contain *aliased* components which introduce artifacts into the reconstruction. The Nyquist conditions are depicted graphically for the one dimensional case in Fig. 7.4-2.

The one dimensional signal l is sampled at rate f_s. It is band limited (as are all real world signals) in the frequency domain with an upper frequency bound of f_B. According to the Nyquist sampling theorem, if a band-limited signal is sampled, the resulting *Fourier* spectrum is made up of the original signal spectrum $|L|$ plus replicates of the original spectrum spaced at integer multiples of the sampling frequency f_s. Diagram (a) in

Table 7.4-1 Digital video application requirements.

Application	Bit rate req.	Distortion req.	Transmission req.	Computational req.	Standards req.
Network video on demand	1.5 Mbps 10 Mbps	High Medium	Internet 100 Mbps lan	MPEG-1 MPEG-2/4	MPEG-1 MPEG-2/4 MPEG-7
Video phone	64 kbps	High distortion	ISDN p x 64	H.261 encoder h.261 decoder	H.261
Desktop multimedia video CDROM	1.5 Mbps	High distortion to medium	PC channel	MPEG-1 decoder	MPEG-1 MPEG-2 MPEG-7
Desktop LAN videoconference	10 Mbps	Medium distortion	Fast ethernet 100 Mbps	Hardware decoders	MPEG-2/4, H.261
Desktop WAN videoconference	1.5 Mbps	High distortion	Ethernet	Hardware decoders	MPEG-1, MPEG-4, H.263
Desktop dial-up videoconference	64 kbps	Very high distortion	POTS and internet	Software decoder	MPEG-4, H.263
Digital satellite television	10 Mbps	Low distortion	Fixed service satellites	MPEG-2 decoder	MPEG-2
HDTV	20 Mbps	Low distortion	12 MHz terrestrial link	MPEG-2 decoder	MPEG-2
DVD	20 Mbps	Low distortion	PC channel	MPEG-2 decoder	MPEG-2

Fig. 7.4-2 depicts the magnitude $|L|$ of the Fourier spectrum for l. The magnitude of the *Fourier* spectrum $|L_s|$ for the sampled signal l_s is shown for two cases. Diagram (b) presents the case where the original signal l can be reconstructed by recovering the central spectral island. Diagram (c) displays the case where the Nyquist sampling criteria has not been met and spectral overlap occurs. The spectral overlap is termed *aliasing* and occurs when $f_s < 2f_B$. When $f_s > 2f_B$, the original signal can be reconstructed by using a low pass digital filter whose pass band is designed to recover $|L|$. These relationships provide a basic framework for the analysis and design of digital signal processing systems.

Two-dimensional or spatial sampling is a simple extension of the one-dimensional case. The Nyquist criteria has to be obeyed in both dimensions, i.e., the sampling rate in the horizontal direction must be two times greater than the upper frequency bound in the horizontal direction, and the sampling rate in the vertical direction must be two times greater than the upper frequency bound in the vertical direction. In practice, spatial sampling grids are square so that an equal number of samples per unit length in each direction are collected. *Charge coupled devices* (CCDs) are typically used to spatially sample analog imagery and video. The sampling grid spacing of these devices is more than sufficient to meet the Nyquist criteria for most resolution and application requirements. The electrical characteristics of CCDs have a greater affect on the image or video quality than its sampling grid size.

Temporal sampling of video signals is accomplished by capturing a spatial or image frame in the time dimension. The temporal samples are captured at a uniform rate of about 60 fields per second for NTSC television and 24 fps for a motion film recording. These sampling rates are significantly less than the spatial sampling rate. The maximum temporal frequency that can be reconstructed according to the Nyquist frequency criteria is 30 Hz in the case of television broadcast. Therefore any rapid intensity change (caused for instance by a moving edge) between two successive frames will cause aliasing because the harmonic frequency content of such a step-like function exceeds the Nyquist frequency. Temporal aliasing of this kind can be greatly mitigated in CCDs by the use of low pass temporal filtering to remove the high frequency content. *Photoconductor storage tubes* are used for recording broadcast television signals. They are analog scanning devices whose electrical characteristics filter the high frequency temporal content and minimize temporal aliasing. Indeed, motion picture film also introduces low pass filtering when capturing image frames. The exposure speed and the response speed of the photo chemical film

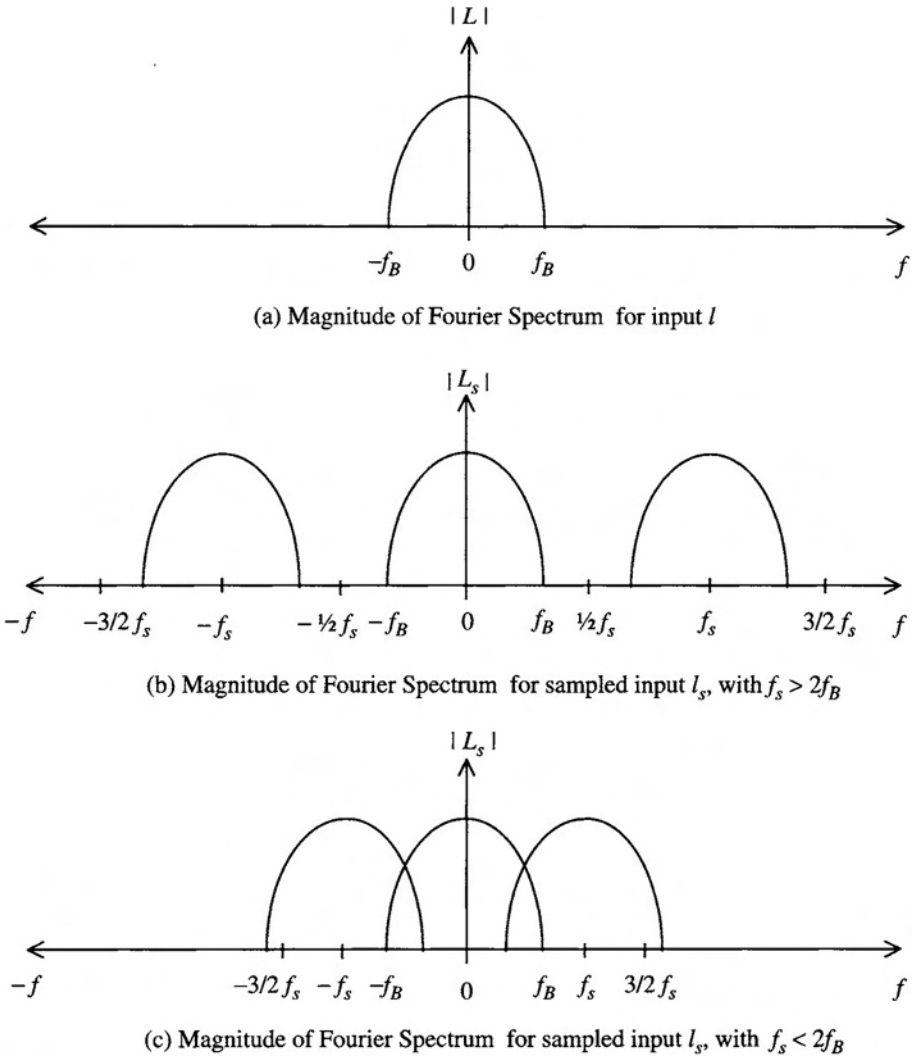


Figure 7.4-2 Nyquist sampling theorem.

combine to mitigate high frequency content and temporal aliasing. These factors cannot completely stop temporal aliasing, so intelligent use of video recording devices is still warranted, e.g., the main reason movie camera panning is done very slowly is to minimize temporal aliasing.

In many cases where fast motions or moving edges are not well resolved due to temporal aliasing, the human visual system will interpolate such motion and provide its own perceived reconstruction. The HVS is very tolerant of temporal aliasing by using its own knowledge of natural motion to provide motion estimation and compensation to the image sequences generated by temporal sampling. The combination of temporal filtering in sampling systems and the mechanisms of human visual perception, reduce the affects of temporal aliasing such that temporal under sampling (sub-Nyquist sampling) is acceptable in the generation of typical image sequences intended for general purpose use.

7.4.4.2 Digital video formats

Sampling is the process used to create the image sequences used for video and digital video applications. Spatial sampling and quantization of a natural video signal digitizes the image plane into a two dimensional set of digital pixels that define a digital image. Temporal sampling of a natural video signal creates a sequence of image frames typically used for motion pictures and television. The combination of spatial and temporal sampling creates a sequence of digital images termed digital video. As described earlier, the digital video signal intensity is defined as $I(i, j, k)$, where $0 \leq i \leq M$, $0 \leq j \leq N$ are the horizontal and vertical spatial coordinates, and $0 \leq k$ is the temporal coordinate.

The standard digital video formats introduced here are used in the broadcast for both analog and digital television, as well as computer video applications. Composite television signal digital broadcasting formats

are included here due to their use in video compression standards, digital broadcasting, and standards format conversion applications. Knowledge of these digital video formats provides background for understanding the international video compression standards developed by the *International Telecommunications Union* (ITU) and the *International Standards Organization* (ISO). These standards contain specific recommendations for use of the digital video formats described here.

Composite television digital video formats are used for the digital broadcasting, SMPTE digital recording, and conversion of television broadcasting formats. Table 7.4-2 contains both analog and digital system parameters for the *National Television Systems Committee* (NTSC), and *Phase Alternating Lines* (PAL) composite broadcast formats.

Component television signal digital video formats have been defined by the *International Consultative Committee for Radio* (CCIR) Recommendation 601. It is based on component video with one luminance (Y) and two color difference signals (C_r and C_b). The raw bit rate for the CCIR 601 format is 162 Mbps. Table 7.4-3 contains important systems parameters of the CCIR 601 digital video studio component recommendation for both NTSC and PAL/SECAM *(sequentiel couleur avec memoire)*.

The ITU Specialist Group (SGXV) has recommended three formats that are used in the ITU H.261, H.263, and ISO MPEG video compression standards. They are the *standard input format* (SIF), *common interchange format* (CIF), and the low bit rate version of CIF called *quarter CIF* (QCIF). Together, these formats describe a comprehensive set of digital video formats that are widely used in current digital video applications. CIF and QCIF support the NTSC and PAL video formats using the same parameters. The SIF format defines different vertical resolution values for NTSC and PAL. The CIF and QCIF formats also support the H.261 modified parameters. The modified parameters are integer multiples of 8 in order to support the 8 x 8 pixel two-dimensional DCT operation. Table 7.4-4 lists this set of digital video standard formats. The modified H.261 parameters are listed in parenthesis.

Table 7.4-2 Digital composite television parameters.

Description	NTSC	PAL
Analog video bandwidth (MHz)	4.2	5.0
Aspect ratio, hor size/vert size	4/3	4/3
Frames per second	29.97	25
Lines per second	525	625
Interlace ratio, fields:frames	2:1	2:1
Subcarrier frequency (MHz)	3.58	4.43
Sampling frequency (MHz)	14.4	17.7
Samples per active line	757	939
Bit rate (Mbps)	114.5	141.9

Table 7.4-3 Digital video component television parameters for CCIR 601.

Description	NTSC	PAL/SECAM
Luminance channel		
Analog video bandwidth (MHz)	5.5	5.5
Sampling frequency (MHz)	13.5	13.5
Samples per active line	710	716
Bit rate (Mbps)	108	108
Two color difference channels		
Analog video bandwidth (MHz)	2.2	2.2
Sampling frequency (MHz)	6.75	6.75
Samples per active line	355	358
Bit rate (Mbps)	54	54

7.4.5 Video compression techniques

Video compression systems are generally comprised of two modes that reduce information redundancy in the spatial and the temporal domains. Spatial compression and quantization operates on a single image block, making use of the local image characteristics to reduce the bit rate. The spatial encoder also includes a variable length coder (VLC) inserted after the quantization stage. The VLC stage generates a lossless encoding of the quantized image block. Temporal domain compression makes use of optical flow models (generally in the form of block-matching motion estimation methods) to identify and mitigate temporal redundancy.

This section presents an overview of some widely accepted encoding techniques used in video compression systems. *Entropy encoders* are lossless encoders that are used in the variable length coding (VLC) stage of a video compression system. They are best used for information sources that are *memoryless* (sources in which each value is independently generated), and try to minimize the bit rate by assigning variable length codes for the input

Table 7.4-4 SIF, CIF, and QCIF digital video formats.

Description	SIF NTSC/PAL	CIF	QCIF
Horizontal resolution (Y) pixels	352	360(352)	180(176)
Vertical resolution (Y) pixels	240/288	288	144
Horizontal resolution (C_r, C_b) pixels	176	180(176)	90(88)
Vertical resolution (C_r, C_b) pixels	120/144	144	72
Bits per pixel (bpp)	8	8	8
Interlace fields:frames	1:1	1:1	1:1
Frame rate (fps)	30	30, 15, 10, 7.5	30, 15, 10, 7.5
Aspect ratio hor size/vert size	4:3	4:3	4:3
Bit rate (Y) Mbps @ 30 fps	20.3	24.9	6.2
Bit rate (U, V) Mbps @ 30 fps	10.1	12.4	3.1

values according to the input *probability density function* (pdf). *Predictive coders* are suited to information sources that have memory, i.e., a source in which each value has a statistical dependency on some number of previous and/or adjacent values. Predictive coders can produce a new source pdf with significantly less statistical variation and entropy than the original. The transformed source can then be fed to a VLC to reduce the bit rate. Entropy and predictive coding are good examples for presenting the basic concepts of statistical coding theory.

Block transformations are the major technique for representing spatial information in a format that is highly conducive to quantization and VLC encoding. Block transforms can provide a coding gain by packing most of the block energy into a fewer number of coefficients. The *quantization* stage of the video encoder is the central factor in determining the rate-distortion characteristics of a video compression system. It quantizes the block transform coefficients according to the bit rate and distortion specifications. Motion compensation takes advantage of the significant information redundancy in the temporal domain by creating current frame predictions based upon block matching motion estimates between the current and previous image frames. Motion compensation generally achieves a significant increase in the video coding efficiency over pure spatial encoding.

7.4.5.1 Entropy and predictive coding

Entropy coding is an excellent starting point in the discussion of coding techniques because it makes use of many of the basic concepts introduced in the discipline of *information theory* or *statistical communications theory* [8]. The discussion of VLC and predictive coders requires the use of *information source* models to lay the statistical foundation for the development of this class of encoder. An information source can be viewed as a process that generates a sequence of symbols from a finite alphabet. Video sources are generated from a sequence of image blocks that are generated from a "pixel" alphabet. The number of possible pixels that can be generated is 2^n, when n is the number of bits per pixel. The order in which the image symbols are generated depends on how the image block is arranged or scanned into a sequence of symbols. Spatial encoders transform the statistical nature of the original image so that the resulting coefficient matrix can be scanned in a manner such that the resulting source or sequence of symbols contains significantly less information content.

Two useful information sources are used in modeling video encoders; the *discrete memoryless source* (DMS), and *Markov* sources. VLC coding is based on the DMS model, and the predictive coders are based on the Markov source models. The DMS is simply a source in which each symbol is generated independently. The symbols are *statistically independent* and the source is completely defined by its symbols/events and the set of probabilities for the occurrence for each symbol, i.e., $E=\{e_1, e_2,\ldots,e_n\}$ and the set $\{p(e_1), p(e_2),\ldots,p(e_n)\}$, where n is the number of symbols in the alphabet. It is useful to introduce the concept of entropy at this point. Entropy is defined as the average information content of

the information source. The information content of a single event or symbol is defined as,

$$I(e_i) = \log\frac{1}{p(e_i)}. \quad (7.4.4)$$

The base of the logarithm is determined by the number of states used to represent the information source. Digital information sources use base 2 in order to define the information content using the number of bits per symbol or bit rate. The entropy of a digital source is further defined as the average information content of the source, i.e.,

$$H(E) = \sum_{i=1}^{n} p(e_i)\log_2\frac{1}{p(e_i)} = -\sum_{i=1}^{n} p(e_i)\log_2 p(e_i) \text{ bits/symbol}. \quad (7.4.5)$$

This relationship suggests that the average number of bits per symbol required to represent the information content of the source is the entropy. The *noiseless source coding theorem* states that a source can be encoded with an average number of bits per source symbol that is arbitrarily close to the source entropy. So called entropy encoders seek to find codes that perform close to the entropy of the source. *Huffman* and *arithmetic* encoders are examples of entropy encoders.

Modified Huffman coding [9] is commonly used in the image and video compression standards. It produces good performing variable length codes without significant computational complexity. The traditional Huffman algorithm is a two step process that first creates a table of source symbol probabilities, and then constructs codewords whose lengths grow according to the decreasing probability of a symbol's occurrence. Modified versions of the traditional algorithm are used in the current generation of image and video encoders. The H.261 encoder uses two sets of static Huffman codewords (one each for AC and DC DCT coefficients). A set of 32 codewords is used for encoding the AC coefficients. The zigzag scanned coefficients are classified according to the zero coefficient run-length and first nonzero coefficient value. A simple table lookup is all that is then required to assign the codeword for each classified pair.

Markov and *random field* source models are well suited to describing the source characteristics of natural images. A Markov source has memory of some number of preceding or adjacent events. In a natural image block, the value of the current pixel is dependent on the values of some of the surrounding pixels because they are part of the same object, texture, contour, etc. This can be modeled as an m-th order Markov source, in which the probability of source symbol e_i depends on the last m source symbols. This dependence is expressed as the probability of occurrence of event e_i conditioned on the occurrence of the last m events, i.e., $p(e_i|e_{i-1}, e_{i-2}, \ldots, e_{i-m})$. The Markov source is made up of all possible n^m states, where n is the number of symbols in the alphabet. Each state contains a set of up to n conditional probabilities for the possible transitions between the current symbol and the next symbol. The *differential pulse code modulation* (DPCM) predictive coder makes use of the Markov source model. DPCM is used in the MPEG-1 and H.261 standards to encode the set of quantized DC coefficients generated by the discrete cosine transforms.

The DPCM predictive encoder modifies the use of the Markov source model considerably in order to reduce its complexity. It does not rely on the actual Markov source statistics at all, and simply creates a linear weighting of the last m symbols (m-th order) to predict the next state. This significantly reduces the complexity of using Markov source prediction at the expense of an increase in the bit rate. DPCM encodes the *differential signal* d between the actual value and the predicted value, i.e., $d = e - \hat{e}$, where the prediction $\hat{e}$ is a linear weighting of m previous values. The resulting differential signal d generally has reduced entropy as compared to the original source. DPCM is used in conjunction with a VLC encoder to reduce the bit rate. The simplicity and entropy reduction capability of DPCM makes it a good choice for use in real-time compression systems. Third order predictors ($m = 3$) have been shown to provide good performance on natural images [10].

7.4.5.2 Block transform coding — the discrete cosine transform

Block transform coding is widely used in image and video compression systems. The transforms used in video encoders are *unitary*, which means that the transform operation has an inverse operation that uniquely reconstructs the original input. The *discrete cosine transform* (DCT), successively operates on 8×8 image blocks, and is used in the H.261, H.263, and MPEG standards. Block transforms make use of the high degree of correlation between adjacent image pixels to provide *energy compaction* or coding gain in the transformed domain. The *block transform coding gain* G_{TC} is defined as the logarithmic ratio of the arithmetic and geometric means of the transformed block variances, i.e.,

$$G_{TC} = 10\log_{10}\left[\frac{\frac{1}{N}\sum_{i=0}^{N-1}\sigma_i^2}{\left(\prod_{i=0}^{N-1}\sigma_i^2\right)^{1/N}}\right], \quad (7.4.6)$$

where the transformed image block is divided into N subbands, and $\sigma^2{}_i$ is the variance of each subband i, for

$0 \le i \le N-1$. G_{TC} also measures the gain of block transform coding over PCM coding. The coding gain generated by a block transform is realized by packing most of the original signal energy content into a small number of transform coefficients. This results in a lossless representation of the original signal that is more suitable for quantization. That is, there may be many transform coefficients containing little or no energy that can be completely eliminated. Spatial transforms should also be orthonormal, i.e., generate uncorrelated coefficients, so that simple scalar quantization can be used to quantize the coefficients independently.

The *Karhunen-Loève transform* (KLT) creates uncorrelated coefficients, and is optimal in the energy packing sense. But the KLT is not widely used in practice. It requires the calculation of the image block covariance matrix so that its unitary orthonormal eigenvector matrix can be used to generate the KLT coefficients. This calculation (for which no fast algorithms exist), and the transmission of the eigenvector matrix is required for every transformed image block.

The DCT is the most widely used block transform for digital image and video encoding. It is an orthonormal transform, and has been found to perform close to the KLT [11] for first-order Markov sources. The DCT is defined on an 8×8 array of pixels,

$$F(u,v) = \frac{1}{4} C_u C_v \sum_{i=0}^{7} \sum_{j=0}^{7} f(i,j) \cos\left(\frac{(2i+1)u\pi}{16}\right) \times \cos\left(\frac{(2j+1)v\pi}{16}\right) \tag{7.4.7}$$

and the inverse IDCT is defined as,

$$f(i,j) = C_u C_v \sum_{u=0}^{7} \sum_{v=0}^{7} F(u,v) \cos\left(\frac{(2i+1)u\pi}{16}\right) \times \cos\left(\frac{(2j+1)v\pi}{16}\right) \tag{7.4.8}$$

with

$$C_u = \frac{1}{\sqrt{2}} \quad \text{for} \quad u = 0, \quad C_u = 1$$

otherwise

$$C_v = \frac{1}{\sqrt{2}} \quad \text{for} \quad v = 0, \quad C_v = 1$$

otherwise

where i and j are the horizontal and vertical indices of the 8×8 spatial array, and u and v are the horizontal and vertical indices of the 8×8 coefficient array. The DCT is the chosen method for image transforms for a couple of important reasons. The DCT has fast $O(n\log n)$ implementations using real calculations. It is even simpler to compute than the DFT because it does not require the use of complex numbers.

The second reason for its success is that the reconstructed input of the *inverse DCT* (IDCT) tends not to produce any significant discontinuities at the block edges. Finite discrete transforms create a reconstructed signal that is periodic. Periodicity in the reconstructed signal can produce discontinuities at the periodic edges of the signal or pixel block. The DCT is not as susceptible to this behavior as the *discrete Fourier transform* (DFT). Since the cosine function is real and even, i.e., $\cos(x) = \cos(-x)$, and the input $F(u, v)$ is real, the IDCT generates a function that is even and periodic in $2n$, where n is the length of the original sequence. On the other hand, the IDFT produces a reconstruction that is periodic in n, but not necessarily even. This phenomenon is illustrated in Figure 7.4-3 for the one dimensional signal $f(i)$.

The original finite sequence $f(i)$ depicted in part (a) is transformed and reconstructed in (b) using the DFT-IDFT transform pairs, and in (c) using the DCT-IDCT transform pairs. The periodicity of the IDFT in (b) is five samples long, and illustrates the discontinuity introduced by the discrete transform. The periodicity of the IDCT in (c) is 10 samples long, due to the evenness of the DCT operation. Discontinuities introduced by the DCT are generally less severe than the DFT. The importance of this property of the DCT is that reconstruction errors and blocking artifacts are less severe in comparison to the DFT. Blocking artifacts are visually striking and occur due to the loss of high frequency components that are either quantized or eliminated from the DCT coefficient array. The DCT minimizes blocking artifacts as compared to the DFT because it doesn't introduce the same level of reconstruction discontinuities at the block edges. Figure 7.4-4 depicts blocking artifacts introduced by gross quantization of the DCT coefficients.

This section ends with an example of the energy packing capability of the DCT. Figure 7.4-5 depicts the DCT transform operation. The original 8×8 image sub-block from the Lena image is displayed in part (a),

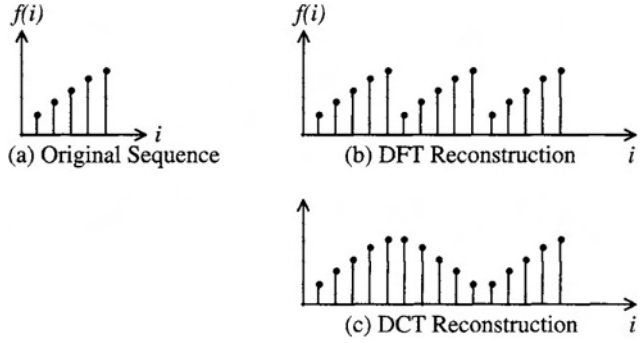


Figure 7.4-3 Reconstruction periodicity of DFT vs DCT.



(a) Original (b) Reconstructed

Figure 7.4-4 Severe blocking artifacts introduced by gross quantization of DCT coefficients.

and the DCT transformed coefficient array is displayed in part (b).

The original image sub-block in (a) contains large values in every position and is not very suitable for spatial compression in this format. The coefficient matrix (b) concentrates most of the signal energy in the top left quadrant. The signal frequency coordinates $(u, v) = (0, 0)$ start at the upper left position. The DC component equals 1255 and contains the vast majority of the signal energy by itself. This dynamic range and concentration of energy should yield a significant reduction in non-zero values and bit rate after the coefficients are quantized.

7.4.5.3 Quantization

The quantization stage of the video encoder creates a lossy representation of the input. The input as discussed earlier should be conditioned with a particular method of quantization in mind. And vice versa, the quantizer should be well matched to the characteristics of the input in order to meet or exceed the rate-distortion performance requirements. As always is the case, the quantizer has an effect on system performance that must be taken under consideration. Simple scalar versus vector quantization implementations can have significant system performance implications.

Scalar and Vector represent the two major types of quantizers. These can be further classified as memoryless or containing memory, and symmetric or nonsymmetric. Scalar quantizers control the values taken by a single variable. The quantizer defined by the MPEG-1 encoder scales the DCT transform coefficients. Vector quantizers operate on multiple variables, i.e., a vector of variables, and become very complex as the number of variables increases. This discussion will introduce the reader to the basic scalar and vector quantizer concepts that are relevant to image and video encoding.

The uniform scalar quantizer is the most fundamental scalar quantizer. It possesses a nonlinear staircase input-output characteristic that divides the input range into output levels of equal size. In order for the quantizer to effectively reduce the bit rate, the number of output values should be much less that the number of input values. The reconstruction values are chosen to be at the midpoint of the output levels. This choice is expected to minimize the reconstruction MSE when the quantization

$$f(i,j) = \begin{bmatrix} 136 & 141 & 143 & 153 & 152 & 154 & 154 & 156 \\ 143 & 150 & 153 & 156 & 160 & 156 & 155 & 155 \\ 149 & 155 & 161 & 163 & 158 & 155 & 156 & 155 \\ 158 & 161 & 162 & 161 & 160 & 158 & 160 & 157 \\ 157 & 161 & 160 & 162 & 161 & 157 & 154 & 155 \\ 160 & 160 & 161 & 160 & 160 & 156 & 156 & 156 \\ 160 & 161 & 160 & 161 & 161 & 157 & 157 & 156 \\ 162 & 162 & 161 & 161 & 162 & 157 & 157 & 157 \end{bmatrix}$$

(a) Original Lena 8×8 image sub-block

$$F(u,v) = \begin{bmatrix} 1255 & -8 & -9 & -6 & 1 & -1 & -3 & 1 \\ -26 & -20 & -5 & 4 & -1 & 1 & 0 & 1 \\ -9 & -5 & 1 & -1 & 0 & 0 & -1 & 0 \\ -6 & -2 & 0 & 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 2 & 0 & -1 & -1 & 0 \\ -2 & 1 & 2 & 0 & 1 & 1 & 0 & -1 \\ -1 & 0 & 0 & -2 & 0 & 0 & 1 & -1 \\ 1 & 0 & -1 & -2 & 0 & 1 & -1 & 0 \end{bmatrix}$$

(b) DCT Coefficients

Figure 7.4-5 8×8 discrete cosine transform.

errors are uniformly distributed. The quantizers specified in the H.261, H.263, MPEG-1, and MPEG-2 video coders are *nearly* uniform. They have constant step sizes except for the larger *dead zone* area (the input range for which the output is zero).

Non-uniform quantization is typically used for non-uniform input distributions, such as natural image sources. The scalar quantizer that produces the minimum MSE for a non-uniform input distribution will have non-uniform steps. As compared to the uniform quantizer, the non-uniform quantizer has increasingly better MSE performance as the number of quantization steps increases. The Lloyd-Max [12] is a scalar quantizer design that utilizes the input distribution to minimize the MSE for a given number of output levels. The Lloyd-Max places the reconstruction levels at the centroids of the adjacent input quantization steps. This minimizes the total absolute error within each quantization step based upon the input distribution.

Vector quantizers decompose the input into a length n vector. An image for instance, can be divided into $M \times N$ blocks of n pixels each, or the image block can be transformed into a block of transform coefficients. The resulting vector is created by scanning the two dimensional block elements into a vector of length n. A vector $\mathbf{X}$ is quantized by choosing a codebook vector representation $\widehat{\mathbf{X}}$ that is its "closest match." The closest match selection can be made by minimizing an error measure, i.e., choose $\widehat{\mathbf{X}} = \widehat{\mathbf{X}}_i$ such that the MSE over all codebook vectors is minimized,

$$\widehat{\mathbf{X}} = \widehat{\mathbf{X}}_{\mathrm{i}} : \min_i \mathrm{MSE}(\mathbf{X}, \widehat{\mathbf{X}}_i) = \min_i \frac{1}{n} \sum_{j=1}^{n} (x_j - \widehat{x}_j)^2 \tag{7.4.9}$$

The index i of the vector $\widehat{\mathbf{X}}_i$ denotes the codebook entry that is used by the receiver to decode the vector. Obviously the complexity of the decoder is much simpler than the encoder. The size of the codebook dictates both the coding efficiency and reconstruction quality. The raw bit rate of a vector quantizer is,

$$\mathrm{bitrate}_{\mathrm{VQ}} = \frac{\log_2 m}{n} \quad \text{bits/pixel}, \tag{7.4.10}$$

where $\log_2 m$ is the number of bits required to transmit the index i of the codebook vector $\widehat{\mathbf{X}}_i$. The codebook construction includes two important issues that are pertinent to the performance of the video coder. The set of vectors that are included in the codebook determine the bit rate and distortion characteristics of the reconstructed image sequence. The codebook size and structure determines the search complexity to find the minimum error solution for Eq. (7.4.9). Two important VQ codebook designs are the *Linde-Buzo-Gray* (LBG) [13] and *tree search VQ* (TSVQ) [14]. The LBG design is based on the Lloyd-Max scalar quantizer algorithm. It is widely used because the system parameters can be generated via the use of an input "training set" instead of the true source statistics. The TSVQ design reduces VQ codebook search time by using m-ary tree structures and searching techniques.

7.4.5.4 Motion compensation and estimation

Motion compensation [15] is a technique created in the 1960s which is used to increase the efficiency of video encoders. Motion compensated video encoders are implemented in three stages. The first stage estimates objective motion (motion estimation) between the previously reconstructed frame and the current frame. The second stage creates the current frame prediction (motion compensation) using the motion estimates and the previously reconstructed frame. The final stage differentially encodes the prediction and the actual current frame as the prediction error. Therefore, the receiver reconstructs the current image only using the VLC encoded motion estimates and the spatially and VLC encoded prediction error.

Motion estimation and compensation are common techniques used to encode the temporal aspect of a video signal. As discussed earlier, block based motion compensation and motion estimation techniques used in video compression systems are capable of the largest reduction in the raw signal bit rate. Typical implementations generally out-perform pure spatial encodings by a factor of three or more. The interframe redundancy contained in the temporal dimension of a digital image sequence accounts for the impressive signal compression capability that can be achieved by video encoders. Interframe redundancy can be simply modeled as static backgrounds and moving foregrounds to illustrate the potential temporal compression that can be realized. Over a short period of time, image sequences can be described as a static background with moving objects in the foreground. If the background does not change between two frames, their difference is zero, and the two background frames can essentially be encoded as one. Therefore the compression ratio increase is proportional to two times the spatial compression achieved in the first frame. In general, unchanging or static backgrounds can realize additive coding gains, i.e.,

$$\text{Static Back ground Coding Gain } \alpha \, \mathrm{N} \bullet (\text{Spatial Compression Ratio of Background Frame}) \tag{7.4.11}$$

where N is the number of static background frames being encoded. Static backgrounds occupy a great deal of the

image area, and are typical of both natural and animated image sequences. Some variation in the background always occurs due to random and systematic fluctuations. This tends to reduce the achievable background coding gain.

Moving foregrounds are modeled as non-rotational rigid objects that move independent of the background. Moving objects can be detected by matching the foreground object between two frames. A perfect match results in zero difference between the two frames. In theory, moving foregrounds can also achieve additive coding gain. In practice, moving objects are subject to occlusion, rotational and non-rigid motion, and illumination variations that reduce the achievable coding gain. Motion compensation systems that make use of motion estimation methods leverage both background and foreground coding gain. They provide pure interframe differential encoding when two backgrounds are static, i.e., the computed motion vector is (0,0). And the motion estimate computed in the case of moving foregrounds generates the minimum distortion prediction.

Motion estimation is an interframe prediction process falling in two general categories; pel-recursive algorithms [16] and block-matching algorithms (BMA) [17]. The pel-recursive methods are very complex and inaccurate, which restrict their use in video encoders. Natural digital image sequences generally display ambiguous object motions that adversely affect the convergence properties of pel-recursive algorithms. This has led to the introduction of *block-matching motion estimation* which is tailored for encoding image sequences. Block-matching motion estimation assumes that the objective motion being predicted is rigid and non-rotational. The block size of the BMA for the MPEG, H.261, and H.263 encoders is defined as 16 × 16 luminance pixels. MPEG-2 also supports 16 × 8 pixel blocks.

BMAs predict the motion of a block of pixels between two frames in an image sequence. The prediction generates a pixel displacement or motion vector whose size is constrained by the search neighborhood. The search neighborhood determines the complexity of the algorithm. The search for the best prediction ends when the best block match is determined within the search neighborhood. The best match can be chosen as the minimum MSE, which for a full search is computed for each block in the search neighborhood, i.e.,

$$\text{Best Match}_{\text{MSE}} = \min_{m,n} \frac{1}{N^2} \sum_{i=1}^{M} \sum_{j=1}^{N} [I^k(i,j) - I^{k-l}(i+m, j+n)]^2, \qquad (7.4.12)$$

where k is the frame index, l is the temporal displacement in frames, M is the number of pixels in the horizontal direction and N is the number of pixels in the vertical direction of the image block, i and j are the pixel indices within the image block, and m and n are the indices of the search neighborhood in the horizontal and vertical directions. Therefore the best match motion vector estimate $MV(m = x, n=y)$ is the pixel displacement between the block $I^k(i,j)$ in frame k, and the best matched block $I^{k-l}(i+x, j+y)$ in the displaced frame $k-l$. The best match is depicted in Fig. 7.4-6.

In cases where the block motion is not uniform or if the scene changes, the motion estimate may in fact increase the bit rate over the corresponding spatial encoding of the block. In the case where the motion estimate is not effective, the video encoder does not use the motion estimate, and encodes the block using the spatial encoder.

The search space size determines the complexity of the motion estimation algorithm. Full search methods are costly and are not generally implemented in real-time video encoders. Fast searching techniques can considerably reduce computational complexity while maintaining good accuracy. These algorithms reduce the search process to a few sequential steps in which each subsequent search direction is based on the results of the current step. The procedures are designed to find local optimal solutions and cannot guarantee selection of the global optimal solution within the search neighborhood. The *logarithmic* search [18] algorithm proceeds in the direction of minimum distortion until the final optimal value is found. Logarithmic searching has been implemented in some MPEG encoders. The *three-step* search [19] is a very simple technique that proceeds along a best match path in three steps in which the search

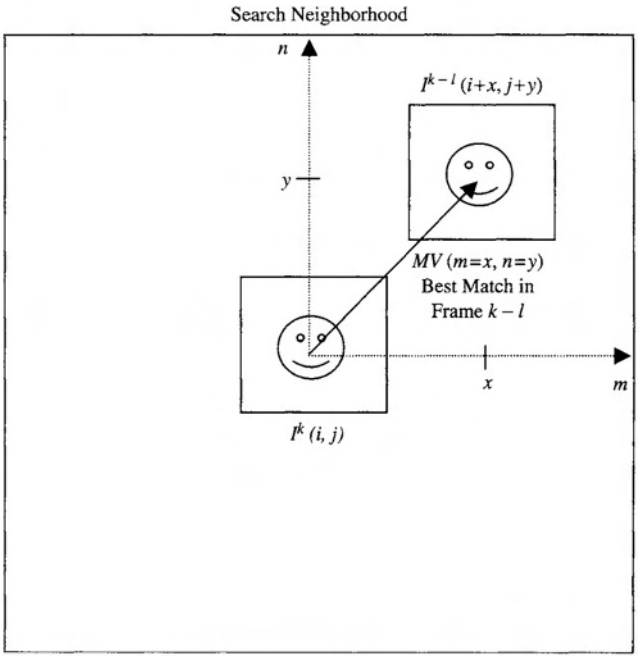


Figure 7.4-6 Best match motion estimate.

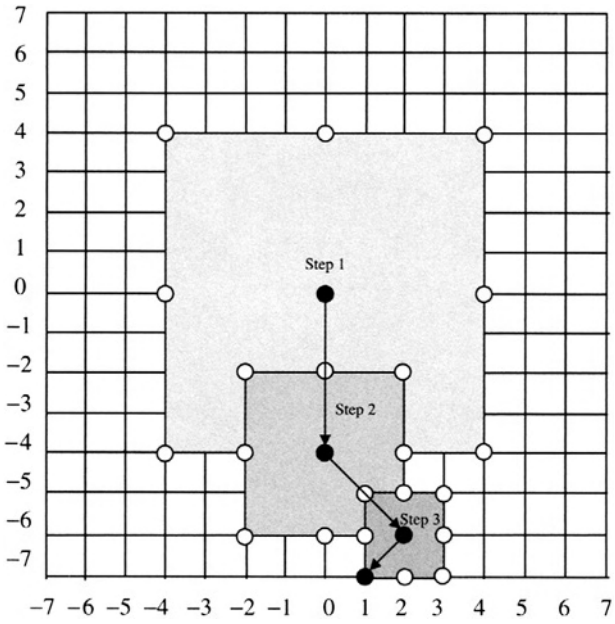


Figure 7.4-7 Three-step search algorithm pictorial.

neighborhood is reduced for each successive step. Figure 7.4-7 depicts the three-step search algorithm.

A 14 × 14 pixel search neighborhood is depicted. The search area sizes for each step are chosen so that the total search neighborhood can be covered in finding the local minimum. The search areas are square. The length of the sides of the search area for step 1 are chosen to be larger than or equal to ½ the length of the range of the search neighborhood (in this example the search area is 8 × 8). The length of the sides are successively reduced by ½ after each of the first two steps are completed. Nine points for each step are compared using the matching criteria. These consist of the central point and eight equally spaced points along the perimeter of the search area. The search area for step 1 is centered on the search neighborhood. The search proceeds in by centering the search area for the next step over the best match from the previous step. The overall best match is the pixel displacement chosen to minimize the matching criteria in step 3. The total number of required comparisons for the three-step algorithm is 25. That represents an 87% reduction in complexity versus the full search method for a 14 × 14 pixel search neighborhood.

7.4.6 Video encoding standards and H.261

The major internationally recognized video compression standards have been developed by the *International Standardization Organization* (ISO), the *International Electrotechnical Commission* (IEC), and the *International Telecommunications Union* (ITU) standards organizations. The *Moving Pictures Experts Group* (MPEG) is a working group operating within ISO and IEC. Since starting its activity in 1988, MPEG has produced ISO/IEC 11172 (MPEG-1, 1992), ISO/IEC 13818 (MPEG-2, 1994), ISO/IEC 14496 (MPEG-4, 1999), and ISO/IEC 15938 (MPEG-7, 2001).

The MPEG-1 specification was motivated by T1 transmission speeds, the CD-ROM, and the multimedia capabilities of the desktop computer. It is intended for video coding up to the rate of 1.5 Mbps, and is composed of five sections: System Configurations, Video Coding, Audio Coding, Compliance Testing, and Software for MPEG-1 Coding. The standard does not specify the actual video coding process, but only the syntax and semantics of the bit stream, and the video generation at the receiver. It does not accommodate interlaced video, and only supports CIF quality format at 25 or 30 fps.

Activity for MPEG-2 was started in 1991. It was targeted for higher bit rates, broadcast video, and a variety of consumer and telecommunications video and audio applications. The syntax and technical contents of the standard were frozen in 1993. It is composed of four parts: Systems, Video, Audio, and Conformance Testing. MPEG-2 was also recommended by the ITU as H.262.

The MPEG-4 project is targeted to enable content developers and users to achieve various forms of interactivity with the audio-visual content of a scene, and to mix synthetic and natural audio and video information in a seamless way. MPEG-4 technology comprises two major parts; a set of coding tools for audiovisual objects, and a syntactic language to describe both the coding tools and the coded objects. From a technical viewpoint, the most notable departure from traditional coding standards is the ability for a receiver to download the description of the syntax used to represent the audio-visual information. The visual information is not restricted to have the format of conventional video, i.e., it may not necessarily be frame-based, but can incorporate audio and/or video foreground and background objects, which can produce significant improvements in both encoder efficiency and functionality.

MPEG-7 is formally named "Multimedia Content Description Interface," and is a common way of describing multimedia content data that is used to access and interpret content by a computer program. Since much of the value of multimedia content can be derived from its accessibility, MPEG-7 strives to define common data access methods to maximize the value of multimedia information regardless of the technologies encompassed by the source and destination, or the specific applications using its services. In order to meet these requirements MPEG-7 has created a hierarchic framework that can handle many levels of description. In

addition other types of descriptive data are defined, such as, coding formats, data access conditions, parental ratings, relevant links, and the overall context. MPEG-7 is made up of three main elements that include description tools, storage and transmission system tools, and a language to define the MPEG-7 syntax. These elements provide the flexibility to meet the stated requirements.

Work on the latest MPEG standard began in 2000. It was started as an extension to MPEG-7 and is known as MPEG-21 (ISO/IEC 21000). It is focused on defining the common content and user access model addressing the vast proliferation of new and old multimedia distribution and reception technologies. MPEG-21 specifically looks to define the technology needed to support users to exchange, access, consume, trade, and otherwise manipulate Digital items in an efficient, transparent and interoperable way. Digital items are defined to be the fundamental unit of distribution and transaction, i.e., content (Web page, picture, movie, etc.).

The precursor to the MPEG video encoding standards development is the H.261 encoder which contains many of the video coding methods and techniques later adopted by MPEG. The ITU Recommendation H.261 was adopted in 1990, and specifies a video encoding standard for videoconferencing and videophone services for transmission over *integrated services digital network* (ISDN) at $p \times 64$ kbps, p=1, …, 30. H.261 describes the video compression methods that were later adopted by the MPEG standards and is presented in the following section. The ITU Experts Group for Very Low Bit-Rate Video Telephony (LBC) has produced the H.263 Recommendation for *Public Switched Telephone Networks* (PSTN), which was finalized in December 1995 [19]. It is an extended version of H.261 supporting bidirectional motion compensation and sub-QCIF formats. The encoder is based on hybrid DPCM/DCT coding and improvements targeted to generate bit rates of less than 64 kbps.

7.4.6.1 The H.261 video encoder

The H.261 recommendation [4] is targeted at the videophone and videoconferencing application market running on connection based ISDN at $p \times 64$ kbps, p = 1,…, 30. It explicitly defines the encoded bit stream syntax and *decoder*, while leaving the encoder design to be compatible with the decoder specification. The video encoder is required to carry a delay of less than 150 msec so that it can operate in real-time bi-directional videoconferencing applications. H.261 is part of a group of related ITU recommendations that define visual telephony systems. This group includes:

1. H.221 — Defines the frame structure for an audiovisual channel supporting 64–1920 kbps.
2. H.230 — Defines frame control signals for audio visual systems.
3. H.242 — Defines audiovisual communications protocol for channels supporting up to 2 Mbps.
4. H.261 — Defines the video encoder/decoder for audiovisual services at $p \times 64$ kbps.
5. H.320 — Defines narrow-band audiovisual terminal equipment for $p \times 64$ kbps transmission.

The H.261 encoder block diagrams are depicted in Fig. 7.4-8 (a) and (b). An H.261 source coder implementation is depicted in (c). The source coder implements the video encoding algorithm that includes the spatial encoder, the quantizer, the temporal prediction encoder, and the VLC. The spatial encoder is defined to use the two dimensional 8×8 pixel block DCT and a nearly uniform scalar quantizer using a possible 31 step sizes to scale the AC and interframe DC coefficients. The resulting quantized coefficient matrix is zigzag scanned into a vector that is VLC coded using a hybrid modified run length and Huffman coder. Motion compensation is optional. Motion estimation is only defined in the forward direction because H.261 is limited to real-time videophone and videoconferencing. The recommendation does not specify the motion estimation algorithm or the conditions for the use of intraframe versus interframe encoding.

The video multiplex coder creates a H.261 bit stream that is based on the data hierarchy described below. The transmission buffer is chosen not to exceed the maximum coding delay of 150 msec, and is used to regulate the transmission bit rate via the coding controller. The transmission coder embeds an error correction code (ECC) into the video bit stream that provides error resilience, error concealment, and video synchronization.

H.261 supports most of the internationally accepted digital video formats. These include, CCIR 601, SIF, CIF, and QCIF. These formats are defined for both NTSC and PAL broadcast signals. The CIF and QCIF formats were adopted in 1984 by H.261 in order to support 525-1ine NTSC and 625-1ine PAL/SECAM video formats. The CIF and QCIF operating parameters can be found in Table 7.4-4. The raw data rate for 30 fps CIF is 37.3 Mbps and 9.35 Mbps for QCIF. CIF is defined for use in channels in which $p \geq 6$ so that the required compression ratio for 30 fps is less than 98:1. CIF and QCIF formats support frame rates of 30, 15, 10, and 7.5 fps, which allows the H.261 encoder to achieve greater coding efficiency by skipping the encoding and transmission of whole frames. H.261 allows 0, 1, 2, 3 or more frames to be skipped between transmitted frames.

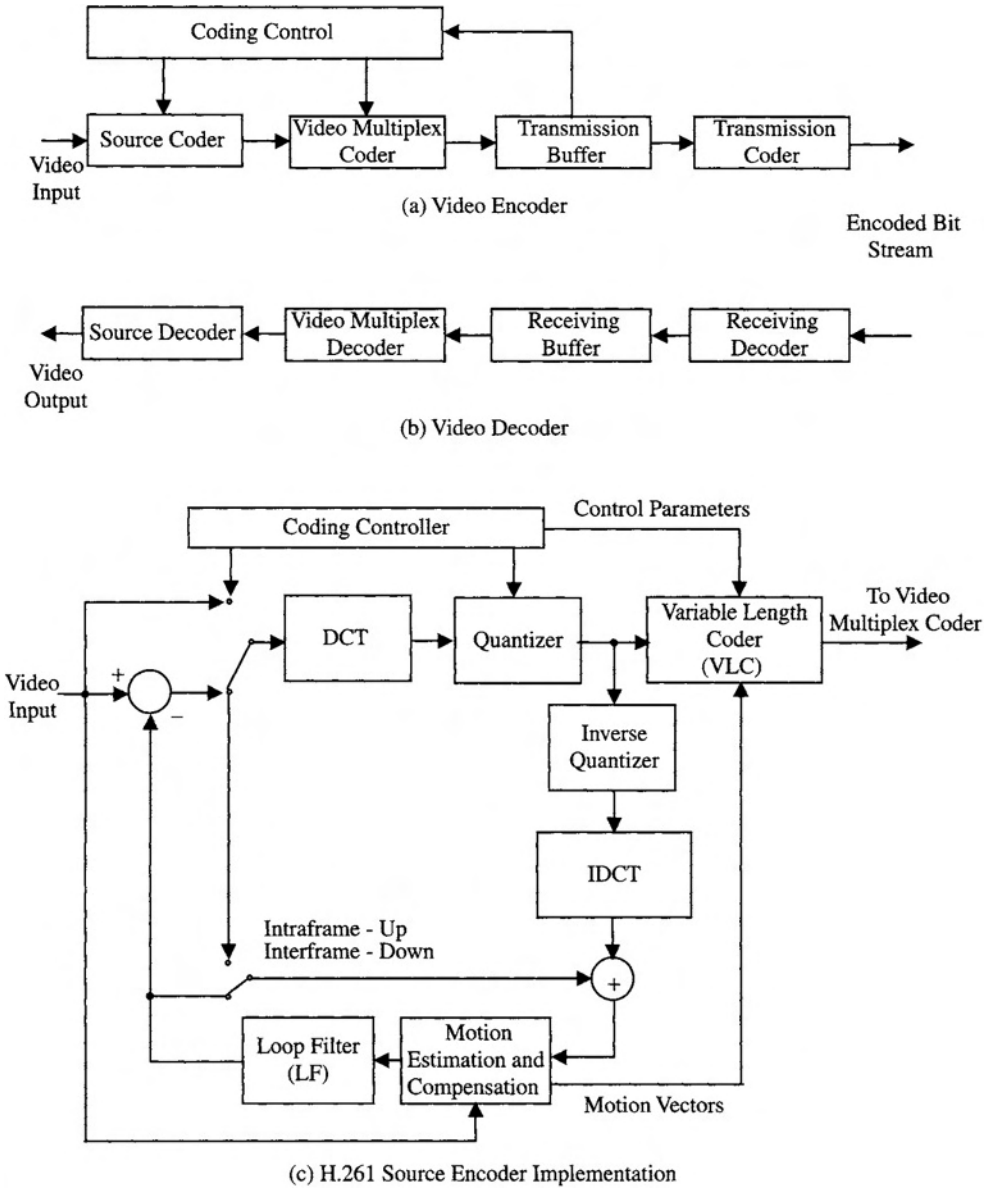


Figure 7.4-8 ITU-T H.261 block diagrams.

H.261 specifies a set of encoder protocols and decoder operations that every compatible system must follow. The H.261 *video multiplex* defines the data structure hierarchy that the decoder can interpret unambiguously. The video data hierarchy defined in H.261 is depicted in Fig. 7.4-9. They are the picture layer, group of block (GOB) layer, macroblock (MB) layer, and the basic (8×8) block layer. Each layer is built from the previous or lower layers, and contains its associated data payload, and header that describes the parameters used to generate the bit stream. The basic 8×8 block is used in intraframe DCT encoding. The macroblock is the smallest unit for selecting intraframe or interframe encoding modes. It is made up of four adjacent 8×8 luminance blocks and two subsampled 8×8 color difference blocks (C_b and C_r as defined in Table 7.4-4) corresponding to the luminance blocks. The GOB is made up of 176×48 pixels (33 macroblocks) and is used to construct the 352×288 pixel CIF or 176×144 pixel QCIF picture layer.

The headers for the GOB and picture layers contain start codes so that the decoder can re-synchronize when errors occur. They also contain other relevant information required to reconstruct the image sequence. The following parameters used in the headers of the data hierarchy complete the H.261 video multiplex.

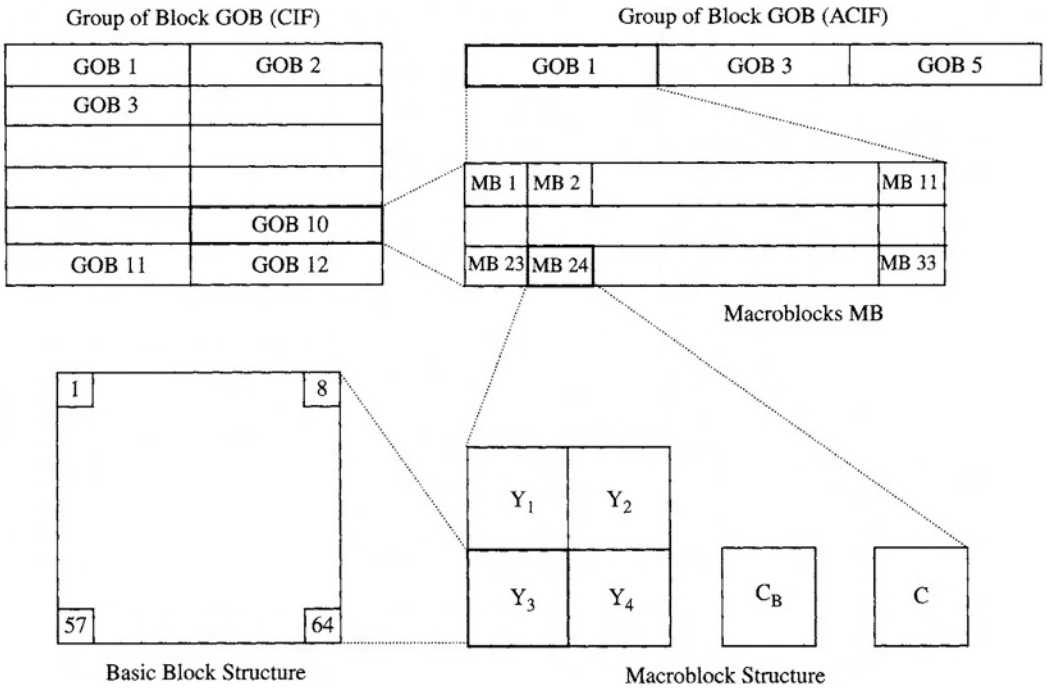


Figure 7.4-9 H.261 block hierarchy.

Picture layer

- Picture start code (PSC), 20-bit synchronization pattern (0000 0000 0000 0001 0000).
- Temporal reference (TR), 5-bit input frame number.
- Type information (PTYPE), indicates source format, CIF=I QCIF=0, and other controls.
- User-inserted bits.

GOB layer

- Group of blocks start code (GBSC), 16-bit synchronization code (0000 0000 0000 0001).
- Group number (GN), 4-bit address representing the 12 GOBs per CIF frame.
- Quantizer information (GQUANT), indicates one of 31 quantizer step sizes to be used in a GOB unless overridden by macroblock MQUANT parameter.
- User-inserted bits.

Macroblock layer

- Macroblock address (MBA), is the position of a macroblock within a GOB.
- Type information (MTYPE), for one of 10 encoding modes used for the macroblock. This includes permutations of intraframe, interframe, motion compensation (MC), and loop filtering (LF). A pre-specified VLC is used to encode these modes.
- Quantizer (MQUANT), 5-bit normalized quantizer step size from 1–31.
- Motion vector data (MVD), up to 11-bit VLC describing the differential displacement.
- Coded block pattern (CBP), up to 9-bit VLC indicating the location of the encoded blocks in the macroblock.

Block layer

- Transform coefficients (TCOEFF), are zigzag scanned and can be 8-bits fixed or up to 13-bit VLC.
- End of block (EOB), symbol.

The H.261 bit stream also specifies transmission synchronization and error code correction using a BCH code [20] that is capable of correcting 2-bit errors in every 511-bit block. It inserts 18 parity bits for every 493 data bits. A synchronization bit is added to every codeword to be able to detect the BCH codeword boundaries. The transmission synchronization and encoding also operates on the audio and control information specified by the ITU H.320 Recommendation.

The H.261 video compression algorithm depicted in Fig. 7.4-8 (c) is specified to operate in intraframe and interframe encoding modes. The intraframe mode provides spatial encoding of the 8 × 8 block, and uses the two-dimensional DCT. Interframe mode encodes the prediction error, with motion compensation being

optional. The prediction error is optionally DCT encoded. Both modes provide options that affect the performance and video quality of the system. The motion estimate method, mode selection criteria, and block transmission criteria are not specified although the ITU has published reference models [21, 22] that make particular implementation recommendations. The coding algorithm used in the ITU-T *Reference Model 8* (RM8) [22] is summarized in three steps, and is followed by an explanation of its important encoding elements.

1. The motion estimator creates a displacement vector for each macroblock. The motion estimator generally operates on the 16×16 pixel luminance macroblock. The displacement vector is an integer value between ± 15, which is the maximum size of the search neighborhood. The motion estimate is scaled by a factor of 2 and applied to the C_r and C_b component macroblocks.
2. The compression mode for each macroblock is selected using a minimum error criteria that is based upon the *displaced macroblock difference* (DMD),

$$DMD(i,j,k) = b(i,j,k) - b(i - d_1, j - d_2, k - 1) \tag{7.4.13}$$

where b is a 16×16 macroblock, i and j are its spatial pixel indices, k is the frame index, and d_1 and d_2 are the pixel displacements of the macroblock in the previous frame. The displacements range from $-15 \leq d_1, d_2 \leq +15$. When d_1 and d_2 are set to zero, the DMD becomes the *macroblock difference* (MD). The compression mode determines the operational encoder elements that are used for the current frame. The H.261 compression modes are depicted in Table 7.4-5.

3. The *video multiplex coder* processes each macroblock to generate the H.261 video bitstream whose elements are discussed above.

There are five basic MTYPE encoding mode decisions that are carried out in step 2. These are,

- Use intraframe or interframe mode?
- Use motion compensation?
- Use a coded block pattern (CBP)?
- Use loop-filtering?
- Change quantization step size MQUANT?

To select the macroblock compression mode, the *variances* (VAR) of the input macroblock, the macroblock Difference (MD) and the displaced macroblock difference (DMD) (as determined by the best motion estimate) are compared as follows,

1. If VAR(DBD) < VAR(MD) then interframe + motion compensation (Inter + MC) coding is selected. In this case, the motion vector data (MVD) is transmitted. Table 7.4-5 indicates that there are three Inter + MC modes that allow for the transmission of the prediction error (DMD) with or without DCT encoding of some or all of the four 8×8 basic blocks.
2. "VAR input" is defined as the variance of the input macroblock. If VAR input < VAR(DMD) and VAR input < VAR(MD) then the intraframe mode (Intra) is selected. Intraframe mode uses DCT encoding of all four 8×8 basic blocks.
3. If VAR(MD) < VAR(DMD) then interframe mode (Inter) is selected. This mode indicates that the motion vector is zero, and that some or all of the 8×8 prediction error (MD) blocks can be DCT encoded.

The transform coefficient coded block pattern (CBP) parameter is used to indicate whether a basic block is reconstructed using the corresponding basic block from the previous frame, or if it is encoded and transmitted. In other words, no basic block encoding is used when the block content does not change significantly. The CPB parameter encodes 63 combinations of the four luminance blocks and two color difference blocks using a variable length code. The conditions for using CBP are not specified in the H.261 recommendation.

Motion compensated blocks can be chosen to be low pass filtered before the prediction error is generated by the feedback loop. This mode is denoted as Inter + MC

Table 7.4-5 H.261 Macroblock video compression modes.

Mode	MQUANT	MVD	CBP	TCOEFF
Intra				✓
Intra	✓			✓
Inter			✓	✓
Inter	✓		✓	✓
Inter + MC		✓		
Inter + MC		✓	✓	✓
Inter + MC	✓	✓	✓	✓
Inter + MC + LF		✓		
Inter + MC + LF		✓	✓	✓
Inter + MC + LF	✓	✓	✓	✓

+ LF in Table 7.4-5. The low pass filter is intended to reduce the quantization noise in the feedback loop, as well as the high frequency noise and artifacts introduced by the motion compensator. H.261 defines loop filtering as optional and recommends a separable two-dimensional spatial filter design which is implemented by cascading two identical one dimensional *finite impulse response* (FIR) filters. The coefficients of the 1D filter are [1, 2, 1] for pixels inside the block, and [0, 1, 0] (no filtering) for pixels on the block boundary.

The MQUANT parameter is controlled by the state of the transmission buffer in order to prevent overflow or underflow conditions. The dynamic range of the DCT macroblock coefficients extends between [−2047,..., 2047]. They are quantized to the range [−127,..., 127] using one of the 31 quantizer step sizes as determined by the GQUANT parameter. The step size is an even integer in the range of [2,..., 62]. GQUANT can be overridden at the Macroblock layer by MQUANT to clip or expand the range prescribed by GQUANT so that the transmission buffer is better utilized. The ITU-T RM8 *liquid level control model* specifies the inspection of 64 Kbit transmission buffers after encoding 11 macroblocks. The step size of the quantizer should be increased (decreasing the bitrate) if the buffer is full, and vice versa, the step size should be decreased (increasing the bitrate) if the buffer is empty. The actual design of the rate control algorithm is not specified.

The DCT macroblock coefficients are subjected to variable thresholding before quantization. The threshold is designed to increase the number of zero valued coefficients, which in turn increases the number of the zero run-lengths and VLC coding efficiency. The ITU-T *Reference Model 8* provides an example thresholding algorithm for the H.261 encoder. Nearly uniform scalar quantization using a dead zone is applied after the thresholding process. All the coefficients in the luminance and chrominance Macroblocks are subjected to the same quantizer except for the intraframe DC coefficient. The intraframe DC coefficient is quantized using a uniform scalar quantizer whose step size is 8. The quantizer decision levels are not specified, but the reconstruction levels are defined in H.261 as follows,

For QUANT odd

$$\text{REC_LEVEL} = \text{QUANT} \times (2 \times \text{COEFF_VALUE} + 1), \quad \text{for COEFF_LEVEL} > 0,$$

$$\text{REC_LEVEL} = \text{QUANT} \times (2 \times \text{COEFF_VALUE} - 1), \quad \text{for COEFF_LEVEL} < 0,$$

For QUANT even

$$\text{REC_LEVEL} = \text{QUANT} \times (2 \times \text{COEFF_VALUE} + 1) - 1, \text{ for COEFF_LEVEL} > 0,$$

$$\text{REC_LEVEL} = \text{QUANT} \times (2 \times \text{COEFF_VALUE} - 1) + 1, \text{ for COEFF_VALUE} < 0.$$

If COEFF_VALUE = 0, then REC_LEVEL = 0, where REC_LEVEL is the reconstruction value, QUANT is ½ the macroblock quantization step size ranging from 1–31, and COEFF_VALUE is the quantized DCT coefficient.

To increase the coding efficiency, lossless variable length coding is applied to the quantized DCT coefficients. The coefficient matrix is scanned in a zig-zag manner in order to maximize the number of zero coefficient run-lengths. The VLC encodes events defined as the combination of a run-length of zero coefficients preceding a nonzero coefficient, and the value of the nonzero coefficient, i.e., EVENT = (RUN, VALUE). The VLC EVENT tables are defined in [4].

7.4.7 Closing remarks

Digital video compression, although only recently becoming an internationally standardized technology, is strongly based upon the information coding technologies researched over the last 40 years. The large variety of bandwidth and video quality requirements for the transmission and storage of digital video information has demanded that a variety of video compression techniques and standards be developed. The major international standards recommended by ISO and the ITU make use of common video coding methods. The generalized digital video encoder introduced in Section 7.4.2, illustrates the spatial and temporal video compression elements that are central to the current MPEG-1, MPEG-2/H.262, MPEG-4, H.261, and H.263 standards that have been developed over the past decade. They address a vast landscape of application requirements, from low to high bitrate environments, as well as stored video and multimedia to real-time videoconferencing and high quality broadcast television.

References

1. ISO/IEC 11172 Information Technology: coding of moving pictures and associated audio for digital storage media at up to about 1.5 Mbit/s, 1993.
2. ISO/IEC JTC1/SC29/WG11, CD 13818: Generic coding of moving pictures and associated audio, 1993.
3. ISO/IEC JTC1/SC29/WG11, CD 14496: Coding of audio-visual objects, 1999.
4. CCITT Recommendation H.261: "Video Codec for Audio Visual Services at $p \times 64$ kbits/s," COM XV-R 37-E, 1990.
5. H. Hseuh-Ming and].W. Woods,"*Handbook of Visual Communications*," Chapter 6, Academic Press Inc., San Diego, CA., 1995.
6. J.W. Woods, "*Subband Image Coding*," Kluwer Academic Publishers, Norwell, MA., 1991.
7. L. Wang and M. Goldberg, "Progressive image transmission using vector quantization on images in pyramid form," *IEEE Trans. Commun.*, 1339–1349, 1989.
8. C.E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, 27, 379–423 and 623–656, July and Oct. 1948.
9. D. Huffman, "A method for the construction of minimum redundancy codes," *Proc. IRE*, 40, 1098–1101, 1952.
10. P.W.]ones and M. Rabbani, "*Digital Image Compression Techniques*," SPIE Optical Engineering Press, Bellingham, WA., 60, 1991.
11. N. Ahmed, T.R. Natarajan, and K.R. Rao, "On image processing and a discrete cosine transform," *IEEE Trans. Comput.*, IT-23, 90–93, Jan. 1974.
12. J.J. Hwang and K.R. Rao, "*Techniques and Standards For Image, Video, and Audio Coding*," Prentice Hall, Upper Saddle River, NJ., 22, 1996.
13. R.M. Gray, "Vector Quantization," *IEEE ASSP Magazine*, IT-1, 4–29, April 1984.
14. W.H. Equitz, "A new vector quantization clustering algorithm," *IEEE Trans. Acous., Speech, Sig. Proc.*, ASSP-37(10), 1568–1575, 1989.
15. B.G. Haskell and J.O. Limb, "Predictive video encoding using measured subjective velocity," U.S. Patent No. 3,632,865, Jan. 1972.
16. A.N. Netravali and J.D. Robbins, "Motion–compensated television coding: Part I," *Bell Syst. Tech. J.*, 58, 631–670, March 1979.
17. J.R. Jain and A.K. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Trans. Commun.*, COM-29, 1799–1808, Dec. 1981.
18. T. Koga et al., "Motion compensated interframe coding for video conferencing," *NTC '81, National Telecommun. Conf.*, G5.3.1–G5.3.5, New Orleans, LA., Nov. 1981.
19. ITU-T SG 15 WP 15/1, *Draft Recommendation H.263* (*Video coding for low bitrate communications*), Document LBC-95-251, Oct 1995.
20. M. Roser et al., "Extrapolation of a MPEG-1 video-coding scheme for low-it-rate applications," *SPIE Video Commun. And PACS for Medical Appl.*, Berlin, Germany, 1977, 180–187, April 1993.
21. CCITT SG XV WP/1/Q4 Specialist Group on Coding for Visual Telephony, *Description of Ref. Model 6* (RM6), Document 396, Oct. 1988.
22. CCITT SG XV WP/1/Q4 Specialist Group on Coding for Visual Telephony, *Description of Ref. Model 8* (RM8), Document 525, June 1989.

Section **Eight**

Appendix

This page intentionally left blank

List of acronyms

Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming

3GPP	Third Generation Partnership Project
AAS	Adaptive Antenna System
ACK	Acknowledgement (in ARQ protocols)
ACK-CH	Acknowledgement Channel (for WiMAX)
ACLR	Adjacent Channel Leakage Ratio
ACTS	Advanced Communications Technology and Services
AGW	Access Gateway (in LTE/SAE)
AM	Acknowledged Mode (RLC configuration)
AMC	Adaptive Modulation and Coding
AMPS	Advanced Mobile Phone System
AMR-WB	Adaptive MultiRate-WideBand
AP	Access Point
ARIB	Association of Radio Industries and Businesses
ARQ	Automatic Repeat-reQuest
ATDMA	Advanced Time Division Mobile Access
ATIS	Alliance for Telecommunications Industry Solutions
AWGN	Additive White Gaussian Noise
BCCH	Broadcast Control Channel
BCH	Broadcast Channel
BE	Best Effort Service
BER	Bit-Error Rate
BLER	Block-Error Rate
BM-SC	Broadcast/Multicast Service Center
BPSK	Binary Phase-Shift Keying
BS	Base Station
BSC	Base Station Controller
BTC	Block Turbo Code
BTS	Base Transceiver Station
CAZAC	Constant Amplitude Zero Auto-Correlation
CC	Convolutional Code
CCSA	China Communications Standards Association
CDF	Cumulative Density Function
CDM	Code-Division Multiplex
CDMA	Code Division Multiple Access
CEPT	European Conference of Postal and Telecommunications Administrations
CN	Core Network
CODIT	Code-Division Testbed
CP	Cyclic Prefix
CPC	Continuous Packet Connectivity
CPICH	Common Pilot Channel
CQI	Channel Quality Indicator
CQICH	Channel Quality Indication Channel (for WiMAX)
CRC	Cyclic Redundancy Check
CS	Circuit Switched
CTC	Convolutional Turbo Code
DCCH	Dedicated Control Channel
DCH	Dedicated Channel
DFE	Decision Feedback Equalization
DFT	Discrete Fourier Transform
DFTS-OFDM	DFT-spread OFDM, see also SC-FDMA
DL	Downlink
DL-SCH	Downlink Shared Channel
DPCCH	Dedicated Physical Control Channel
DPCH	Dedicated Physical Channel
DPDCH	Dedicated Physical Data Channel
DRX	Discontinuous Reception
DTCH	Dedicated Traffic Channel
DTX	Discontinuous Transmission
D-TxAA	Dual Transmit-Diversity Adaptive Array
E-AGCH	E-DCH Absolute Grant Channel
E-DCH	Enhanced Dedicated Channel
EDGE	Enhanced Data rates for GSM Evolution and Enhanced Data rates for Global Evolution
E-DPCCH	E-DCH Dedicated Physical Control Channel
E-DPDCH	E-DCH Dedicated Physical Data Channel
E-HICH	E-DCH Hybrid Indicator Channel
eNodeB	E-UTRAN NodeB
EPC	Evolved Packet Core
E-RGCH	E-DCH Relative Grant Channel
ErtPS	Extended Real-Time Polling Service
E-TFC	E-DCH Transport Format Combination
E-TFCI	E-DCH Transport Format Combination Index

Communications Engineering Desk Reference; ISBN: 9780123746481
Copyright © 2007, Erik Dahlman, Stefan Parkvall, Johan Skold and Per Beming. Published by Elsevier Ltd 2001. All rights reserved

ETSI	European Telecommunications Standards Institute
E-UTRA	Evolved UTRA
E-UTRAN	Evolved UTRAN
EV-DO	Evolution-Data Optimized (of CDMA2000 1x)
EV-DV	Evolution-Data and Voice (of CDMA2000 1x)
FACH	Forward Access Channel
FBSS	Fast Base Station Switching
FCC	Federal Communications Commission
FCH	Frame Control Header (for WiMAX)
FDD	Frequency Division Duplex
FDM	Frequency-Division Multiplex
FDMA	Frequency-Division Multiple Access
F-DPCH	Fractional DPCH
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FIR	Finite Impulse Response
F-OSICH	Forward link Other Sector Indication Channel (for IEEE 802.20)
FPLMTS	Future Public Land Mobile Telecommunications Systems
FRAMES	Future Radio Wideband Multiple Access Systems
FTP	File Transfer Protocol
FUSC	Fully Used Subcarriers (for WiMAX)
GERAN	GSM EDGE RAN
GGSN	Gateway GPRS Support Node
GPRS	General Packet Radio Services
GPS	Global Positioning System
G-RAKE	Generalized RAKE
GSM	Global System for Mobile communications
HARQ	Hybrid ARQ
H-FDD	Half-duplex FDD
HHO	Hard Handover
HLR	Home Location Register
HSDPA	High-Speed Downlink Packet Access
HS-DPCCH	High-Speed Dedicated Physical Control Channel
HS-DSCH	High-Speed Downlink Shared Channel
HSPA	High-Speed Packet Access
HS-PDSCH	High-Speed Physical Downlink Shared Channel
HSS	Home Subscriber Server
HS-SCCH	High-Speed Shared Control Channel
HSUPA	High-Speed Uplink Packet Access
IDFT	Inverse DFT
IEEE	Institute of Electrical and Electronics Engineers
IFDMA	Interleaved FDMA
IFFT	Inverse FFT
IMS	IP Multimedia Subsystem
IMT-2000	International Mobile Telecommunications 2000
IP	Internet Protocol
IPv4	IP version 4
IPv6	IP version 6
IR	Incremental Redundancy
IRC	Interference Rejection Combining
ISDN	Integrated Services Digital Network
ITU	International Telecommunications Union
ITU-R	International Telecommunications Union-Radiocommunications Sector
Iu	The interface used for communication between the RNC and the core network.
Iub	The interface used for communication between the NodeB and the RNC
Iur	The interface used for communication between different RNCs
J-TACS	Japanese Total Access Communication System
LAN	Local Area Network
LDPC	Low Density Parity Check Code
LMMSE	Linear Minimum Mean-Square Error
LTE	Long-Term Evolution
MAC	Medium Access Control
MAN	Metropolitan Area Network
MAP	Map message (for WiMAX)
MBFDD	Mobile Broadband FDD (for IEEE 802.20)
MBMS	Multimedia Broadcast/Multicast Service
MBS	Multicast and Broadcast Service
MBSFN	Multicast Broadcast Single Frequency Network
MBTDD	Mobile Broadband TDD (for IEEE 802.20)
MBWA	Mobile Broadband Wireless Access
MCCH	MBMS Control Channel
MCE	MBMS Coordination Entity
MCH	Multicast Channel
MCS	Modulation and Coding Scheme
MDHO	Macro-Diversity Handover
MICH	MBMS Indicator Channel
MIMO	Multiple-Input Multiple-Output
ML	Maximum Likelihood
MLD	Maximum Likelihood Detection
MMS	Multimedia Messaging Service
MMSE	Minimum Mean Square Error
MRC	Maximum Ratio Combining
MSC	Mobile Switching Center
MSCH	MBMS Scheduling Channel
MTCH	MBMS Traffic Channel
NAK	Negative Acknowledgement (in ARQ protocols)
NMT	Nordisk MobilTelefon (Nordic Mobile Telephony)
NodeB	NodeB, a logical node handling transmission/reception in multiple cells. Commonly, but not necessarily, corresponding to a base station.
nrTPS	Non-Real-Time Polling Service
OFDM	Orthogonal Frequency-Division Multiplexing
OFDMA	Orthogonal Frequency-Division Multiple Access
OOK	On-off keying
OVSF	Orthogonal Variable Spreading Factor
PAN	Personal Area Network
PAPR	Peak-to-Average Power Ratio
PAR	Peak-to-Average Ratio (same as PAPR)
PARC	Per-Antenna Rate Control
PCCH	Paging Control Channel
PCG	Project Coordination Group (in 3GPP)
PCH	Paging Channel

PCI	Pre-coding Control Indication
PDC	Personal Digital Cellular
PDCCH	Physical Downlink Control Channel
PDCP	Packet-Data Convergence Protocol
PDSCH	Physical Downlink Shared Channel
PDU	Protocol Data Unit
PF	Proportional Fair (a type of scheduler)
PHY	Physical layer
PoC	Push to Talk over Cellular
PS	Packet Switched
PSK	Phase Shift Keying
PSTN	Public Switched Telephone Network
PUSC	Partially Used Subcarriers (for WiMAX)
QAM	Quadrature Amplitude Modulation
QoS	Quality-of-Service
QPSK	Quadrature Phase-Shift Keying
RAB	Radio Access Bearer
RACE	Research and development in Advanced Communications technologies in Europe
RAN	Radio Access Network
RB	Resource Block
RBS	Radio Base Station
RF	Radio Frequency
RLC	Radio Link Protocol
RNC	Radio Network Controller
RNTI	Radio-Network Temporary Identifier
ROHC	Robust Header Compression
RR	Round Robin (a type of scheduler)
RRC	Radio Resource Control
RRM	Radio Resource Management
RS	Reference Symbol
RSN	Retransmission Sequence Number
RSPC	IMT-2000 Radio Interface Specifications
RTP	Real Time Protocol
rtPS	Real-Time Polling Service
RTWP	Received Total Wideband Power
RV	Redundancy Version
S1	The interface between eNodeB and AGW.
SA	System Aspects
SAE	System Architecture Evolution
S-CCPCH	Secondary Common Control Physical Channel
SC-FDMA	Single-Carrier FDMA
SDMA	Spatial Division Multiple Access
SDO	Standards Developing Organisation
SDU	Service Data Unit
SF	Spreading Factor
SFBC	Space Frequency Block Coding
SFN	Single-Frequency Network or System Frame Number (in 3GPP)
SFTD	Space Frequency Time Diversity
SGSN	Serving GPRS Support Node
SIC	Successive Interference Combining
SIM	Subscriber Identity Module
SINR	Signal-to-Interference-and-Noise Ratio
SIR	Signal-to-Interference Ratio
SMS	Short Message Service
SNR	Signal-to-Noise Ratio
SOHO	Soft Handover
SRNS	Serving Radio Network Subsystem
STBC	Space-Time Block Coding
STC	Space-Time Coding
STTD	Space-Time Transmit Diversity
TACS	Total Access Communication System
TCP	Transmission Control Protocol
TD-CDMA	Time Division-Code Division Multiple Access
TDD	Time Division Duplex
TDM	Time Division Multiplexing
TDMA	Time Division Multiple Access
TD-SCDMA	Time Division-Synchronous Code Division Multiple Access
TF	Transport Format
TFC	Transport Format Combination
TFCI	Transport Format Combination Index
TIA	Telecommunications Industry Association
TM	Transparent Mode (RLC configuration)
TR	Technical Report
TrCH	Transport Channel
TS	Technical Specification
TSG	Technical Specification Group
TSN	Transmission Sequence Number
TTA	Telecommunications Technology Association
TTC	Telecommunications Technology Committee
TTI	Transmission Time Interval
UE	User Equipment, the 3GPP name for the mobile terminal
UGS	Unsolicited Grant Service
UL	Uplink
UL-SCH	Uplink Shared Channel
UM	Unacknowledged Mode (RLC configuration)
UMTS	Universal Mobile Telecommunications System
USIM	UMTS SIM
US-TDMA	U.S. TDMA standard
UTRA	Universal Terrestrial Radio Access
UTRAN	Universal Terrestrial Radio Access Network
Uu	The interface used for communication between the NodeB and the UE.
WAN	Wide Area Network
WARC	World Administrative Radio Congress
WCDMA	Wideband Code Division Multiple Access
WG	Working Group
WiMAX	Worldwide Interoperability for Microwave Access
WLAN	Wireless Local Area Network
VoIP	Voice-over-IP
WP8F	Working Party 8F
WRC	World Radiocommunication Conference
X2	The interface between eNodeBs.
ZC	Zadoff-Chu
ZF	Zero Forcing
ZTCC	Zero Tailed Convolutional Code

This page intentionally left blank

Index

2nd channel interference, 38
3Com Corp., 175
3G evolution:
  driving forces, 267
    cost and performance, 269–70
    design services, 269
    Internet and IP technology, 268
    service needs, wide spectrum of, 269
    technology advancements, 267–8
    telephony services, 268–9
  history and background, 259–61
  radio access network evolution, 270–2
  spectrum for, 264–6
  standardization:
    3GPP, 262–3
    IMT-2000 activities, in ITU, 263–4
    process, 261–2
  system architecture evolution, 272
3GPP TSG RAN, 262, 270
4:1:1 YCbCr format, 473, 474
4:2:0 YCbCr format, 474, 475
4:2:2 YCbCr format, 473
4:4:4 YCbCr format, 473–4, 475, 476
4-PSK, 212
4-Quadrature Amplitude Modulation (4-QAM), 216, 217
16QAM modulation, 216–18, 277
50 Ω line, 52
64-QAM, 216–18
75 Ω line, 52
140 Ω line, 52
300 Ω line, 52
480i and 480p systems:
  interlaced analog component video, 484–5
  interlaced analog composite video, 484
  interlaced digital component video, 486–8
  progressive analog component video, 485–6
  progressive digital component video, 488
  SIF and QSIF, 488
480i video systems, 484
576i and 576p systems:
  interlaced analog component video, 488–9
  interlaced analog composite video, 488
  interlaced digital component video, 490
  progressive analog component video, 489–90
  progressive digital component video, 490–2
576i video systems, 484
600 Ω line, 52
720p systems
  progressive analog component video, 492
  progressive digital component video, 492–3, 494
802.11 WLANs, 386
  802.11 (In) Security, 401–3
  802.11 alphabet soup, 392–4
  802.11 architecture, 387, 388
  802.11 classic direct-sequence PHY, 389–92, 393
  802.11a, 121
  802.11a PHY, 396–9, 400
  802.11g PHY, 399–400
  MAC and CSMA/CA, 387–9
  Wi-Fi PHY (802.11b), 394–6
1080i and 1080p systems:
  interlaced analog component video, 493, 495
  interlaced digital component video, 495–6
  progressive analog component video, 495
  progressive digital component video, 496–7
1152i format, 493, 495

AAL2, 168, 174
Absorbing Boundary Condition (AbBC), 317
Abstraction, 158
Accelerated Retroactive Decoding (ARD), 228
'Acceptance angle', 29
Access point (AP), 387, 388, 389, 415
ACK method, 232
Acknowledgment (ACK) packet, 224
Adaptive Clocking, 167
Adaptive Modulation and Coding (AMC), 296
Adaptive radios, 118, 122–3
Additive white Gaussian noise (AWGN) channel, 205
Adjacent channel selectivity, 35
Admission-control, 195
Admission control agents, 196
ADPCM (Adaptive Differential Pulse Code Modulation), 167
Advanced Communication Technologies and Services (ACTS), 261
Advanced networking, 10
Advanced radio classes, properties of, 124
Advanced retransmission schemes, 302
Alternating Direction Implicit (ADI) method, 317
Amplified aerial distribution systems, using amplifiers, 32–3
Amplified antenna distribution systems, 32
Amplitude modulation (AM), 17–18
AMPS (Advanced Mobile Phone Service), 259
AN/ARC-164 HaveQuick II, 117
AN/ARC-210, 119
AN/ARC-220, 117
AN/PRC-117F, 119
AN/PRC-138 Falcon, 120
AN/PRC-6725, 118
AN/VRC-99, 117
AN/WSC-3, 117
Analog-to-digital converters (ADCs), 94
Analog video signals, sampling of, 506–8
Analog vs. digital video, 463

Antenna/aerials, 23, 147–8
  beamwidth, 148
  directive gain, 148
  distribution systems, 29
    amplified aerial distribution systems, using amplifiers, 32–3
    amplified antenna distribution systems, 32
    balanced and unbalanced systems, 29–30
    matching network, 32
    multi-point antenna distribution systems, 30–2
  folded dipole antenna, 25
  half-wave dipole antenna, 24–5
  isotropic, 146
  monopole/vertical rod antenna, 25
  multi-loop antennas, 26–8
  radiating resistance, 24
  single loop antennas, 25–6
Antenna architectures, 341
Antenna arrays:
  Yagi-Uda array, 28–9
Application class, 106
Application programming interface (API), 107–8, 109
Application softwares, 109
  Binary Runtime Environment for Wireless (BREW), 110
  Java, 109–10
  Python, 110
Application-specific integrated circuit (ASIC), 76
ApplicationFactory, 106
Ardis, 8
Artificial intelligence (AI) techniques, 125–6
Aspect-oriented programming (AOP), 101
Association of Radio Industries and Businesses (ARIB), 261
Asynchronous connectionless link (ACL), 426, 428, 429
Asynchronous transfer mode (ATM), 167, 199, 403–4
ATDMA project (Advanced TDMA Mobile Access), 260
ATM adaptation layer 1 (AAL1), 168
ATM trunking, using AAL2, 174
  AAL2 common part sub-layer (CPS), 174
  AAL2 service-specific convergence sub-layer (SSCS), 174
  non-switched trunking, 174–5
  switched trunking, 174
Audio and video compression, 465–6
Audio–video synchronization, 242–3
Authority and Format Identifier (AFI), 169
Automatic gain control (AGC), 76
Automatic Repeat Request (ARQ), 201, 302
  hybrid FEC/ARQ protocols, 220
    type-I hybrid ARQ protocols, 220
    type-II hybrid-ARQ protocols, 220–21
  pure ARQ protocols, 219
    go-back-N ARQ, 219
    selective-repeat ARQ, 219–20
    stop-and-wait ARQ, 219
Automatic retransmission on request, 242
Available bit rate (ABR), 168
Avaya Inc., 175
Aware radios, 118, 120

Back off interval, 416
Balanced antenna, 29
Balanced line, 29–30
Balanced/unbalanced transformer, 30
Bandwidth (bw), definition of, 18
Bandwidth-limited operation, 274
Barker sequence, 390, 419, 420
Baseball rule, 166
Baseband processing deployment, 98–9
Baseband processor engines:
  digital signal processors, 96–7
  field-programmable gate arrays, 97–8
  general purpose processors, 96
Basic service set (BSS), 387, 414, 415
Basic telephony, 166–7
Beacon instant, 428
Beam Launching (BL) models, 325
Beam tracking models, *see* Beam Launching (BL) models
Bell Operating Companies (BOCs), 160
Best-effort traffic (BBE service), 195
Binary erasure channel (BEC), 204
Binary Phase Shift Keying (BPSK), 216, 390, 391
Binary Runtime Environment for Wireless (BREW), 112
Binary symmetric channel (BSC), 204
Biometric sensors, 125–6
Bit error rate (BER), 123, 204, 392, 395, 405
Bit interleaving, 162
Block layer, 519–21
Block-matching algorithms (BMA), 515
Block-matching motion estimation, 504, 515
Block transform coding, 511–13
Block transform coding gain, 512
Bluetooth, 8, 11, 425
  active mode, 427
  error correction and encryption, 428–9
  hold mode, 427
  inquiry and paging, 429–30
  packet format, 428
  park mode, 428
  sniff mode, 427
Bluetooth and Wi-Fi coexistence, improving, 435
  adaptive frequency hopping, 435
  packet fragmentation, 436
  power control, 435
Bluetooth personal area network, 7
Bluetooth Special Interest Group (SIG), 404, 425
BNC (baby N connector), 46
Branch-line coupler, 65–6
Broadband Yagi arrays, 28
Broadcast channels, 157–8
BT.601, 483, 486, 490
BT.1358, 488, 490
Buffering delay, 171, 240, 241
Burst range, 198
Byte interleaving, 162

Call setup process, example of, 237
CANCEL method, 232
Carrier sense multiple access with collision avoidance (CSMA/CA), 388, 416, 417
Carrier sensing, 385

CCK modulation, 421, 422
CDMA2000, 86–9
Cell construction delay, 171, 174
Cell Delay Variation (CDV), 171, 173
Cell-switching networks, 176
CELP (Code Excited Linear Prediction), 171
Channel-associated signaling (CAS), 167, 169, 171
Channel capacity, 203–4, 273
Channel code, 201, 203
Channel codeword, 201, 204, 205
Channel coding, 203–4, 287, 288
Channel coding and error control, for bit errors and packet losses, 205
  convolutional codes, 210–14
  interleaving, 214
  linear block codes, 206
    CRC codes, 208
    digital fountain codes, 209
    irregular repeat-accumulate (IRA) codes, 209
    lattice codes, 209
    LDPC codes, 208–9
    Reed-Solomon codes, 208
    tornado codes, 209
  turbo codes, 214–15
Channel decoder, 201
Channel-dependent scheduling, 295
  acquiring on channel-state information, 300–1
  downlink scheduling, 296–9
  and link adaptation, 300
  and traffic behavior, 301–2
  uplink scheduling, 299–300
Channel encoder, 201
Channel errors, 201, 506
Channel measurements:
  device performance and reliability, of channel structure estimates, 345–7
  measurement scenarios, 339–40
  and parameter estimation, 341–5
Channel protection fundamentals, 201
  channel coding and error control, for bit errors and packet losses, 205
    convolutional codes, 210–14
    interleaving, 214
    linear block codes, 206–10
    turbo codes, 214–15
  hierarchical modulation, 216–18
  hybrid ARQ protocols, 220–21
  pure ARQ protocols, 219–20
  Shannon's source and channel theorems, 202
    channel coding, 203–5
    extensions, 205
    source coding, 202–3
    source-channel coding, 204–5
  sounding techniques, 340–1
Channelization codes, 309
Characteristic impedance (Z_0), of transmission lines, 20, 42
  by measurement, 50
    open-circuit measurement, 51
    short-circuit measurement, 51–2
  coaxial line characteristic impedance, 46
  microstrip line characteristic impedance, 46–8
  from primary electrical parameters, 48–50
  twin parallel wire characteristic impedance, 46
Charge coupled devices (CCD), 507, 508
Chase combining, 303, 304
Chromaticity diagram, 478–80
Circuit emulation services (CES), 173–4
Cisco Systems Inc., 175
Class-based queuing (CBQ), 196–7
Class of Service (CoS) vs. Quality of Service (QoS), 194
Clear to send (CTS) packet, 129, 389, 417
Client-side techniques, 241
Coaxial line, 44–5
  characteristic impedance, 46
Code division multiple access (CDMA), 71, 74, 86, 88, 390
Code division multiple access (CDMA) based cellular system, 120
Code Domain Multiplexing (CDM), 296, 298
Codec selection, 192
Coded block pattern (CBP), 520
Codewords, 202
Coding efficiency, 502, 518, 521
Coding gain, 304, 395, 396, 512, 515
CODIT project (Code Division Test bed), 260
Cognitive radio (CR), technologies required for, 69, 91, 115, 121–2
  adaptive radios, 120, 122–3
  aware radios, 120, 122
  biometrics, 125–6
  funding and research, 130
    authentication applications, 136–7
    cognitive geolocation applications, 131–32
    dynamic spectrum access and spectrum awareness, 132–5
    rendezvous problem, 135–6
  geolocation, 125
  radio capabilities and properties, comparison of, 125
  radio flexibility and capability, 117–20
  software technology, 127
    AI techniques, 127–8
    networking protocols, 129
    policy engines, 127
    signal processing, 128–9
    software communications architecture, 129–30
  spatial awareness or situational awareness, 126–7
  spectrum awareness and potential for sublease or borrow, 130
  spectrum awareness/frequency occupancy, 125
  time, 126
  timeline for, 137
Cognitive waveform development, 113–14
Collaborative signal and information processing (CSIP), 253
Collision detection, 385
Colour spaces:
  chromaticity diagram, 478–80
  constant luminance problem, 482
  gamma correction, 480
    current systems, 481–2
    early NTSC systems, 481
    early PAL and SECAM systems, 483

Colour spaces: (*continued*)
non-CRT displays, 482
HLS (hue, lightness, saturation) color space, 477
HSI (hue, saturation, intensity) color space, 477–8
HSV (hue, saturation, value) color space, 477
non-RGB color space considerations, 480
PhotoYCC color space:
PhotoYCC to RGB, 477
RGB to PhotoYCC, 475, 477
RGB color space:
scRGB, 470
sRGB, 469
xvYCC color space, 474
YCbCr color space:
4:4:4 YCbCr format, 473–4, 476
RGB-YCbCr equations, 471–2
RGB-YCbCr equations, 472–3
YIQ color space, 471
YUV color space, 470–1
Combined modulation methods, 19–20
Commercial cable impedances, 52
Commercial off-the-shelf (COTS) hardware, 120
Committed Information Rate (CIR), 198, 199
Common channel signaling (CCS), 167, 171
Common interchange format (CIF), 509, 518
Common Object Request Broker Architecture (CORBA), 102, 103
Common part sub-layer (CPS), of AAL2, 172
Common Pilot Channel (CPICH), 309
Complementary code keying (CCK), 394, 420
Complex instruction set computers (CISCs), 96
Component-based programming (CBP), 100–1
Compressed Real-time Transport Protocol (cRTP), 191, 195
Compression algorithms, 169
Compression ratio, 502, 504, 515
Congestion–distortion optimized (CoDiO) scheduling, 228–9
Connected Limited Device Configuration (CLDC), 110
Connectionless switching, 155–6
Constant bit rate (CBR), 168, 170, 173
Continuous wave (CW), 144
Convolutional codes, 210–14
Coplanar waveguides, 44
COST Action, 259, 316
Coupled lines, 45
Couplers, 64
branch-line coupler, 65–6
'rat-race' coupler, 66–7
ring coupler, 66
Cramer-Rao Lower Bound (CRLB), 345–6
CRC codes, 208
Cross-polarisation Power Ratios (XPRs), 360
cRTP (Compressed Real Time Protocol), 197
CSEL (Combat Survivor/Evader Locator), 120
CSRC, 235–6
Cutback method, 453
Cutoff wavelength, 445, 452
Cyclic-prefix insertion, 284–5
Cyclic redundancy check (CRC), 203, 302, 308, 428
DARPA Agent Markup Language (DAML), 125, 128
Data and voice traffic, differences in, 153
data traffic characteristics, 156
broadcast channels, 157–8
data networks, 159–60
layering and routing, 159
protocol layers, 158–9
protocol stacks, 159
packet switching vs. circuit switching, 151
connectionless switching, 153–4
frame switching, 152–3
packet switching, 153
voice traffic characteristics, 160
convergence technologies, 162–4
data, addition of, 161–62
TDM architecture, 162
Data country code (DCC), 169
Data networks, 157, 159–60, 175, 176, 181, 189
Data traffic characteristics, 156
broadcast channels, 157–8
data networks, 159–60
layering and routing, 159
protocol layers, 158
layered communications systems, 158
levels of abstraction, 158
peer communications, 158
protocol stacks, 159
Decibel, use of, 148–9
Dedicated inquiry access codes (DIAC), 429
Dedicated Physical Channel (DPCH), 309
Dedicated Physical Control Channel (DPCCH), 310
Dedicated Physical Data Channel (DPDCH), 309
Defense Advanced Research Projects Agency (DARPA), 128
Delay since last SR (DLSR), 236
Demodulation circuits, *see* Demodulators
Demodulators, 16
Demultiplexer, 162, 167
Dependent packet delays, 227–8
Deterministic propagation modelling, 320
electromagnetic models, 324
input database issues, 320–4
electromagnetic parameters determination, 321–2
geometric database simplification, 323–4
mixed database handling, 322–3
ray models, 324–7
advanced techniques for, 327–32
applications, 333–7
simplified or hybrid models, 337–9
Device Configuration Descriptor (DCD), 107
Device Package Descriptor (DPD), 107
DeviceManager, 105–6, 107
Differential equation methods:
finite difference time domain models, 317–18
parabolic equation (PE) methods, 318–19
Differential pulse code modulation (DPCM), 171, 511
Differential Services Code Point (DSCP), 192
Differentiated Services (Diff-Serv), 185, 187, 194
Digital component video background, 483
coding ranges, 483

EDTV sample rate selection, 484
HDTV sample rate selection, 484
SDTV sample rate selection, 484
Digital Data Service (DDS), 160
Digital European cordless telephone (DECT), 123
Digital fountain codes, 209
Digital media adapters, 466, 467
Digital modulation, 17
Digital processing, in SDR, 83–6
Digital signal processing, 71, 76–7, 126
Digital signal processors, 71, 93, 96–7, 126–7 *see also* Fixed function DSP; General purpose DSP
Digital television set-top box, 467
Digital video signals and formats, 506
analog video signals, sampling of, 506–8
digital video formats, 508–9, 510
Direct-sequence spread-spectrum (DSSS) radio link, 131, 386
Directive gain of an antenna, 146
Discard Eligibility (DE) bit, 175
Discrete cosine transform (DCT), 503, 511–13
Discrete Fourier transform (DFT), 512, 513
Discrete memoryless source (DMS), 200, 201, 203, 510, 511
Dispersion, in transmission systems, 43
Displaced macroblock difference (DMD), 520
'Distortionless transmission', 49
Distributed coordination function interframe space (DIFS), 388, 416
Distributed interframe space (DIFS), 388, 389
DMR (Digital Modular Radio), 121
Dodecahedral antenna, 314
Domain Profile, 107
Domestic Yagi arrays, 28
Double superheterodyne receivers, 39
Downlink scheduling, 296–9
Duplex, 190
Duty cycle, 437–8
DVD players, 466–7
Dynamic Bandwidth Circuit Emulation Services (DBCES), 172
Dynamic frequency selection (DFS), 394, 425

E&M interface, 166
Echo, 191
EDTV sample rate selection, 484
Effective Aperture Distribution Function (EADF) concept, 346–7
Electric field strength, 145
Electrical transmission lines, types of, 44
coaxial and strip lines, 44–5
coupled lines, 45
microstrip and slot lines, 45
twin lines, 45
waveguides and coplanar waveguides, 44
Electromagnetic spectrum, 20, 21
Electromagnetic Vector Sensor (EVS), 355
Electromagnetic waves, 141
EM theory and diffuse scattering:
differential equation methods:
finite difference time domain models, 317–18
parabolic equation methods, 318–19
integral methods:
full moment methods, 314–15
modal expansion models, 316
physical optics models, 315–16
ray methods:
Gaussian beams, 319
geometrical optics models, 319–20
Empirical path-loss models, for micro-cells, 349–50
Encoding delay, 171, 238
Energy density, 145
Enhanced-definition video, 465
Enhanced messaging service (EMS), 8
Entropy and predictive coding, 510–11
Entropy coding, 503, 510–11
Entropy encoders, 510
Entropy of stationary source, 203
Enveloped carrier wave, 16
EPLRS (Enhanced Position Location Reporting System), 119
Equivalent isotropic radiated power (EIRP), 148
Equivalent radiated power (ERP), 148
Ericsson, 8
Error correction codes (ECC), 506, 517
Error-detection techniques, 239
Error-resilience techniques, 239
ESPRIT technique, 342
Ethernet frame, 387
European Broadcasting Union (EBU), 483
ExecutableDevice, 105
Extended service set (ESS), 387, 415
Extended-SuperFrames, 167
Extensible Markup Language (XML), 106, 107
Extrinsic information, 215

Fading prediction, 367–70
Fast Fourier transform (FFT), 399
and OFDM implementation, 282–4
Federal Communications Commission (FCC), 3, 141
FHS (frequency hop synchronization), 429
Fiber optic cable designs, 453, 454, 455–6
Fiber optic connectors:
mechanical properties, 455, 456–7
optical properties, 457, 458
types, 454, 455, 456
Field-programmable gate arrays (FPGAs), 76, 97–8
Finite difference time domain models, 317–18
Finite impulse response (FIR) filters, 521
First-generation (1G) cellular systems, 71
Fisher Information Matrix (FIM), 347
Fixed function DSP, 83–5
Fixed location systems, 4, 5
free space optical (FSO), 6
point to multipoint, 6
point-to-point RF, 5–6
very small aperture terminal (VSAT) systems, 6
Folded dipole antenna, 25, 52
Foreign exchange office (FXO) interface, 166
Foreign exchange service (FXS) interface, 166
Foreman sequence, 225, 228

Forward error correction (FEC), 94, 199, 239, 386, 423, 424, 428, 429
'Forward gain', 28–9
Frame relay, 198, 199
Frame switching, 152–3
Free space optical (FSO) communications, 6
Free-space radiation, 20–3
Frequency, 145
  and wavelength, 141–2
Frequency attenuation, 43
Frequency diversity, 287
Frequency Division Duplex (FDD), 264, 265–6
Frequency division multiplexing, 156, 167
Frequency-domain model, of OFDM, 285–6
Frequency-Domain Multiplexing, 296, 298
Frequency-hopping, 120, 390
Frequency-hopping radio link, 386
Frequency-hopping spread spectrum (FHSS), 413, 414, 418
Frequency interleaving, 287, 288
Frequency modulation (FM), 17, 18–19
'Front to back ratio', 29
Full-duplex circuit, 162
Full function device (FFD), 432
Future Public Land Mobile Telecommunications Systems (FPLMTS), 260
Future Radio Wideband Multiple Access System (FRAMES), 261

Galois fields, 204
Gamma correction, 480
  current systems, 481–2
  early NTSC systems, 481
  early PAL and SECAM systems, 483
  non-CRT displays, 482
Gateway, 155
Gaussian beams, 319
Gaussian frequency-shift keying, *see* Gaussian minimum-shift keying (GMSK)
Gaussian minimum-shift keying (GMSK), 405
General inquiry access code (GIAC), 429
General Packet Radio Services (GPRS), 260
General purpose DSP, 69, 76–7, 85–6
General purpose processors (GPPs), 96
Geographic routing, 254
Geolocation, 123, 129–30
Geometrical optics (GO) theory, 324–5
Geo-synchronous earth orbit (GEO) satellites, 6
Gigabit Ethernet, 199
Glare, 164
Global positioning system (GPS), 123
Global system for mobile communication (GSM), 71, 74
Globalstar, 7
GNURadio, 101, 118
Goal-based agent, 126
Greedy filling, 298, 300
Ground start, 166
Group of block (GOB) layer, 518–19

H.261 video encoder, 517–21
H.261 video multiplex, 518–19
Half-duplex circuit, 162
Half-wave dipole antenna, 24–5
Hamming distance, 204, 206, 210
Hardware controller, in SCA, 105
Hazeltine, 70
Heterodyning, 142
Hierarchical modulation, 201, 216–18
High data rates, in mobile communication, 273
  fundamental constraints, 273–6
    in interference-limited scenarios, 275–6
    in noise-limited scenarios, 274–5
  higher-order modulation, 276–8
  wider bandwidth including multi-carrier transmission, 278–80
High data security, 425
High-definition video, 465
High-priority (HP) bit stream, 216, 217, 218
High resolution and bandwidth (HDTV), 505
  RGB-YCbCr equations, 472–3
  sample rate selection, 484
High resolution parameter estimation techniques, 342, 344
High Speed Downlink Packet Access (HSDPA), 263, 305
High-Speed Packet Access (HSPA), 305
  evolution, *see* HSPA Evolution
Higher-order modulation, 276–8
HiperLAN and HiperLAN, 2, 403–4
HIPERLAN/2, 424
HLS (hue, lightness, saturation) color space, 477
HomeRF Working Group, 414
Hopping clock, 405
Horizontal sync, 464
HSI (hue, saturation, intensity) color space, 477–8, 479
HSPA evolution, 273–4
HSV (hue, saturation, value) color space, 477, 478
Human visual system (HVS), 503, 504
Hybrid ARQ, 220, 295, 302–4
  type-I hybrid ARQ protocols, 220
  type-II hybrid ARQ protocols, 220–21
Hybrid ring coupler, 66
Hybrid-Space SAGE (HS-SAGE) algorithm, 344
HyperText Markup Language (HTML), 107

IEEE 802.1 p/Q, 194
Image channel interference, 38
Impedance relations, in transmission lines, 63–4
Improperly terminated line, 54
IMT-2000, in ITU, 263–4
Incident waves, 42
Incremental Redundancy (IR), 303–4
Independent, identically distributed (iid) model, 229
Infinite fields, 206
Infrared Data Association (IrDA), 6
Infrared link, 386
Infrared signals, 6
Initialisation and Search Improved SAGE (ISIS) algorithm, 343
Inmarsat, 7
Input impedance, of low loss transmission lines, 63
Institute of Electrical and Electronic Engineers (IEEE) 802.16, 6, 384
Inter-Access Point Protocol (IAPP), 387

Inter-cell interference, 275
Interexchange Carriers (IECs), 160
Interface definition language (IDL), 102
Interferometric method, 453, 455
Interframe mode, 502, 503
Interlaced vs. progressive displays, 464–5
International code designator (ICD), 167
International Consultative Committee for Radio (CCIR), 509
International Telecommunications Union (ITU), 509
Internet and IP technology, 268
Internet access, 8
Internet Protocol (IP), 10, 154
Internet Protocol (IP)-based real-time communication, *see* Real-time communication
Internet protocol voice and video telephony and teleconferencing, *see* Real-time communication
Interworking functions (IWFs), 173
Intra-cell interference, 275
Intraframe mode, 502, 503
Inverse DCT (IDCT), 512
Inverse discrete Fourier transform (IDFT), 283, 423, 424
Inverse fast Fourier transform (IFFT) processing:
  and OFDM implementation, 282–4
INVITE method, 234
Ionising and non-ionising radiations, 144
IP header, 193
IP network protocols, to support voice, 193
  Class of Service (CoS) vs. Quality of Service (QoS), 194
  class-based queuing (CBQ), 196–7
  differential services, 194
  IEEE 802.1 p/Q, 194
  prioritized queuing, 196
  priority queuing, 196
  Resource Reservation Protocol (RSVP), 195–6
  UDP ports, 194
  VLANs, 194
  weighted fair queuing (WFQ), 196
IP softphone, 190–1
IrDA (Infrared Data Association), 418
Iridium, 7
Irregular repeat-accumulate (IRA) codes, 209
Isotropic antenna, 24, 148
Isotropic light source, 23
Iterative Sensitivity Adjustment (ISA) algorithm, 225
ITU, IMT-2000 in, 263–4
ITU BT.709, 484
ITU BT.1358, 484
ITU-R BT.709, 495, 496
ITU-T Reference Model 8 (RM8), 521
IXCs, *see* Interexchange Carriers (IECs)

Jaguar PRC-116, 119
Java, 109–10
Java 2 Enterprise Edition (J2EE), 110
Java 2 Micro Edition (J2ME), 110
Java 2 Standard Edition (J2SE), 109
Java VM (JVM), 109
Jitter, 182, 188
Joint Pictures Expert Group (JPEG), 503
JTRS (Joint Tactical Radio System), 121, 127
JTT (Joint Tactical Terminal), 119

Karhunen-Lodve transform (KLT), 512
Kirchhoff models, 315, 330

LANs to PANs operation, 404
  802.15.3 task group, 406–7
  Bluetooth, 404–6
  UWB PANs, 407–11
Last SR timestamp (LSR), 238
Lattice codes, 209–10
Layering and routing, 159
LDCELP (Low Delay CELP), 17
LDPC codes, 208–9
Leprechaun, *see* AN/PRC-6725
Levels of abstraction, 158
LifeCycle class, 103–4
Light emitting diode (LED), 447
Light propagation, in optical fiber,
  multimode fiber, 446–7
  optical coupling, 447–8
  rays and electromagnetic mode theory, 443
  single-mode fiber, 443–6
Linde-Buzo-Gray (LBG), 514
Line characteristic impedances, of transmission lines, 45
  coaxial line characteristic impedance, 46
  microstrip line characteristic impedance, 46–8
  twin parallel wire characteristic impedance, 46
Linear block codes, 206
  CRC codes, 208
  digital fountain codes, 209
  irregular repeat-accumulate (IRA) codes, 209
  lattice codes, 209
  LDPC codes, 208–9
  Reed–Solomon codes, 208
  Tornado codes, 209
Linear finite state machine (LFSM), 210
Linear programming (LP), 100
Link adaptation, 295–6
  and channel-dependent scheduling, 300
List Viterbi algorithm (LVA), 212
Listed directory number (LDN), 163
Lloyd-Max, 514
LoadableDevice class, 105
Local Access and Transport Areas (LATAs), 158
Local Area Network (LAN), 153–4, 155, 157, 383–4
Local area systems, 4, 6–7
  Bluetooth personal area networks, 7
  Infrared Data Association (IrDA), 6
  infrared signals, 6
  wireless local area networks, 7
Local Exchange Carriers (LECs), 160
Local multipoint distribution system (LMDS), 6
Log Periodic Yagis, 28
Logical Channel Number (LCN), 155
Logical link control (LLC) layer, 415
Long-Term Evolution (LTE), 259, 270
  drivers and philosophy, 271

Long-term fading, 352–4
LoS formula, 339
Loss-concealment techniques, 241
Lossless source coding, 202
Lossy source coding, 203
Low earth orbit (LEO) satellite systems, 7
Low-noise amplifier (LNA), 75
Low-priority (LP) bit stream, 216, 217–18
Lower medium access control (MAC) layer., 384
LST-5E, 118
LubyTransform (LT) codes, 207
Lucent Technologies, 175

Machine-readable policy-controlled radios, 121
Macro diversity, *see* Soft handover
Macroblock (MB) layer, 518, 519
Magnetic field strength, 147
Managed Multiservice Networking (MMN), 175
Markov source model, 511
M-ary biorthogonal keying (MBOK), 408, 409
Matched line, 54
Matching network, 32
Max-C/I (maximum rate) scheduling, 297
Max-C/I scheduler, 301
Maximum-distance separable (MDS) codes, 207
Maximum Transmission Unit (MTU) size, 197
MBITR (MultiBand Intra/Inter Team Radio), 118, 120
MBMMR (MultiBand, MultiMode Radio), 120
Mean absolute error (MAE) criteria, 502
Mean Opinion Score (MOS), 192
Mean power, 145
Mean square error (MSE), 502
Media gateway, 163
Media transport and control protocols, 236
  call setup process, example of, 239
  real-time control protocols, 238
  real-time transport protocols, 236–8
  video payload format, 238–9
Medium Access Control (MAC) layer, 94, 414
  and CSMA/CA, 387–9
  in WCDMA, 308
Medium earth orbit (MEO) satellite systems, 7
Memory management unit (MMU), 96
Messaging, 8
Metropolitan area networks (MANs), 6, 157, 159–60, 384
Microstrip line, 45
  characteristic impedance, 46–8
Microwaves, 142
MIDS (Multifunction Information Distribution System), 118
MIMO (Multiple-Input Multiple-Output), 92, 275, 305, 355
Minimum distance decoding, 202
MIPS (millions of instructions per second), 171
Mismatched loss, 57
ML technique, 342
Mobile communication, high data rates in, *see* High data rates, in mobile communication
Mobile Internet services vs. Internet services, 268
Mobile IP, 10, 268
Mobile radio channel characterisation:
  fading prediction, 367–70
  long-term fading, 352–4
  narrow band short-term fading, 354–6
  path loss and building penetration, 348–52
  statistical processing of measurements, 347–8
  temporal and angular dispersion, 359–67
  wideband short-term fading, 356–9
Mobile satellite services, 7
Mobitex, 8
Modal expansion model, 316
Mode field diameter (MFD), 445
Model-based reflex agent, 126
Modified Huffman coding, 511
Modulated carrier wave, *see* Enveloped carrier wave
Modulation:
  definition, 16
  methods, 16
    amplitude modulation (AM), 17–18
    combined modulation methods, 19–20
    frequency modulation (FM), 18–19
    phase modulation (PM), 19
Modulators, definition of, 16
Monopole/vertical rod antenna, 25
Moore's law, 267
Mother code, 213
Motient, 8
Motion compensation and estimation, 503–4, 514–16
Motion Pictures Expert Group (MPEG), 501, 516, 517
Motorola, 7, 70
Motorola 56001, 97
MPEG-1, 516
MPEG-2, 495, 516
MPEG-4, 495, 516–17
MPEG-7, 517
MPEG-21, 517
MSK (minimum shift keying), 432
Multiband radio, 70
Multi-carrier/multi-channel radio, 70
Multi-carrier receiver, 73–4, 75
Multi-carrier transmission, 279–80
Multi-cell broadcast/multicast transmission, 292–3
Multichannel multipoint distribution service (MMDS), 6
Multicore systems, 99
Multilayer switches, 185
Multi-loop antennas, 26–8
Multimedia Broadcast Multicast Services (MBMS), 305
Multimedia messaging service (MMS), 8
Multimode (MM) fiber, 445, 446, 447, 449
Multimode radio, 70
Multipath fading, 4
Multiple deadlines, 226–7
Multiplexer, 160, 165
Multi-point antenna distribution systems, 30–2
Multi-rate radio, 70
MUSIC algorithm, 342

Narrowband fading prediction, 368–9
Narrowband short-term fading, 354–6
National Science Foundation (NSF), 128–9

National Telecommunications and Information Administration (NTIA), 129
National Television Systems Committee (NTSC), 505, 518
Near-field method, 452, 453
Negative acknowledgment (NACK), 219, 227
Network-adaptive media transport, 223
  congestion–distortion optimized (CoDiO) scheduling, 230–41
  dependent packet delays, 229–30
  multiple deadlines, 228–9
  rate–distortion optimized (RaDiO) streaming, 224
    RaDiO framework, 224–6
    receiver-driven streaming, 226, 227
  rich acknowledgments, 226–8
Network Address Translation (NAT), 198
Network allocation vector (NAV), 388
Network delay, 240
Network infrastructure, 165
  basic telephony, 166–7
  time division multiplexing (TDM), 167
  voice networking, 165–6
  Voice over ATM (VoATM), 167
    addressing, 169
    ATM trunking, using AAL2, 174–5
    benefit, 170
    circuit emulation services (CES), 173–4
    and delay, 170
    routing, 169–70
    signaling, 169
    standards and specifications, 172–3
    technical challenges, 170–72
  Voice over Frame Relay (VoFR), 175
    equipping network, for VoFR, 175–8
    voice switching requirements, 178–9
  Voice over IP (VoIP), 179–87
Network layer prioritization, 185
Network packet loss, 191–2
Networking protocols, 127
Next-generation network, 154
NeXt Generation (XG) radio development program, 127–8
NMT (Nordic Mobile Telephony) system, 259
NodeBs, 306
Noise figure, 36
  for ADC, 83
Noise spectral density (NSD), 82
Noiseless source coding theorem, 511
Non-CRT displays, 482
Non-ionising radiations, 144
Non-orthogonal multiple-access scheme, 299
Non-RGB color space considerations, 480
Non-switched trunking, 174–5
Nortel Networks Corp., 175
NTDR (Near-Term Digital Radio) system, 120
'Number of elements', 28
Numerical aperture (NA), 447
Nyquist sampling theorem, 506–7
Nyquist theorem, 166

Object-oriented programming (OOP), 100
Obstruction probability, 337
Omni-Tracs, 7
Ontology Inference Layer (OIL), 125
Open-circuit impedance, 50, 51
Open System Interconnection Reference Model (OSI/RM), 415, 416
Open Systems Interconnect (OSI), 108, 384
Optical fiber:
  bragg gratings, 457–9
  cable designs, 453, 454, 455–6
  characterization:
    attenuation, 449–50
    dispersion, 450–2
    materials, 448–9
    measurement techniques, 452, 453, 455
    mechanical properties, 451, 452
  connector:
    mechanical properties, 455, 456–7
    optical properties, 457, 458
    types, 454, 455, 456
  light propagation,
    multimode fiber, 446–7
    optical coupling, 447–8
    rays and electromagnetic mode theory, 443
    single-mode fiber, 443–6
OPTIONS method, 234
Orthogonal frequency division multiplexing (OFDM), 398, 410
  basic parameters, selection of:
    cyclic-prefix length, 289–90
    number of subcarriers, 288–9
    subcarrier spacing, 288
  basic principles, 281–2
  channel coding, importance of, 287
  channel estimation and reference symbols, 286–7
  cyclic-prefix insertion, 284–5
  demodulation, 282
  frequency-domain model, 285–6
  implementation using IFFT/FFT processing, 282–4
  instantaneous transmission power, variations in, 290
  multi-cell broadcast/multicast transmission, 292–3
  as user-multiplexing and multiple-access scheme, 290–2
Orthogonal multiple-access scheme, 299
Orthogonal Variable Spreading Factor (OVSF) codes, *see* Channelization codes
Outdoor-to-indoor propagation, 326

Packet binary convolutional coding (PBCC), 394, 420–1
Packet Data Convergence Protocol (PDCP), 306, 308
Packet delay, 181, 190
Packet loss, 190–91
Packet mis-order, 191
Packet-switching networks, 176
Packet switching vs. circuit switching, 153
  connectionless switching, 155–6
  frame switching, 154–5
  packet switching, 155
Packetization delay, 171, 240
Paging, 7
Parabolic equation (PE) methods, 318–19
Patch antennas, 341

PCM (Pulse Code Modulation), 171
PCS, 417
Peak Cell Rate (PCR), 173
Peak pulse power density, 147
Peak signal-to-reconstruction noise (PSNR), 502
Peer programs, 159
Pel-recursive algorithms, 515
Periodic time (T), 43
Permanent Virtual Circuits (PVCs), 155, 168, 169, 177
Personal area network (PAN), 7, 11, 383, 432
Phase delay, 43
Phase modulation (PM), 17, 19
Photoconductor storage tubes, 508
PhotoYCC color space:
  PhotoYCC to RGB, 477
  RGB to PhotoYCC, 475, 477
Physical layer, 417–24, 425
  802.11a and OFDM, 423–4
  802.11b, 420–3
  DSSS PHY, 419–20, 421
  FHSS PHY, 418–19
  IEEE 802.11 basic, 418
  infrared PHY, 418
  of WCDMA, 308–10
Piconet, 425
Picture layer, 518, 519
Plain old telephone service (POTS), 166
Plane wave, definition of, 21
Platform configuration tool (PCT), 164
Platform-specific analysis, 98–9
PLRS (Position Location Reporting System), 117
Point of Presence (POP), 160
Point-to-point (PPP) circuits, 199
Polarisation of electromagnetic wave, 21
Policy-based management tool, 184
Port class, 104
PortSupplier, 104
Positive acknowledgment (ACK), 217
Post Office Code Standardization Advisory (POCSAG)
    numeric paging code, 3
Power, definition of, 147
Power flux density, 147
Power-limited operation, 274
Predictive coders, 510
Primary Rate Interface, 165
Prioritization/CoS (class of service), 184–5
Prioritized queuing, 196
Private Branch Exchange (PBX), 161
Private network-to-network interface (PNNI), 169
Probabilistic decoding, 209
Progressive transmission, 505
Progressive vs. interlaced displays, 464–5
Propagation constant (γ), of transmission lines, 59–62
Propagation modelling, 313
  and channel characterisation
    fading prediction, 367–70
    long-term fading, 352–4
    narrow band short-term fading, 354–6
    path loss and building penetration, 348–52
    statistical processing of measurements, 347–8
    temporal and angular dispersion, 359–67
    wideband short-term fading, 356–9
  channel measurements
    channel sounding techniques, 340–1
    device performance and reliability of channel structure estimates, 345–7
    measurement scenarios, 339–40
    and parameter estimation, 341–5
  deterministic propagation modelling, 320
    electromagnetic models, 324
    input database issues, 320–4
    ray models, 324–37
    simplified or hybrid models, 337–9
  EM theory and diffuse scattering:
    differential equation methods, 317–19
    integral methods, 314–16
    ray methods, 319–20
Propagation time delay, 43
Properly terminated transmission line, 54
Properties Descriptor, 107
PropertySet class, 102–3
Proportional-fair (PF) scheduler, 301
Protocol layers, 156
  layered communications systems, 158
  levels of abstraction, 158
  peer communications, 158
Protocol stacks, 159
Protocol suites, 159
PSDU (PHY service data unit), 432
Pseudo-random bit sequence, 419
Public switched telephone network (PSTN), 163
Pulse code modulated (PCM), 162, 168
Pulse duty factor (DF), 147
Pulse propagation, in transmission line, 52–3
Pulse repetition frequency (p.r.f.), 147
Pure ARQ protocols, 219
  go-back-N ARQ, 219
  selective-repeat ARQ, 219–20
  stop-and-wait ARQ, 219
Python, 110, 113

QPSK modulation, 276, 277
QSIF, 488
Quadrature amplitude modulation (QAM), 20
Quadrature Phase Shift Keying (QPSK), 216, 423
Quality of service (QoS), 10, 168, 193, 424
  audio-video synchronization, 242–3
  available bandwidth, adapting to, 242
  combating losses, 241
    active techniques, 241–42
    client-side techniques, 241
  minimizing latency, 239–41
Quarter CIF (QCIF), 509, 518

Racal, 25, 119
Radiating resistance, 24
Radio access network (RAN) evolution, 270–2
Radio capabilities and properties, comparison of, 122, 123

Radio channel characterisation, *see* Mobile radio channel characterisation
Radio communication systems, 15
  antenna arrays,
    Yagi-Uda array, 28–9
  antenna distribution systems, 29
    amplified aerial distribution systems, using amplifiers, 32–3
    amplified antenna distribution systems, 32
    balanced and unbalanced systems, 29–30
    matching network, 32
    multi-point antenna distribution systems, 30–2
  antennas and aerials, 23
    folded dipole antenna, 25
    half-wave dipole antenna, 24–5
    monopole/vertical rod antenna, 25
    multi-loop antennas, 26–8
    radiating resistance, 24
    single loop antennas, 25–6
  modulation and demodulation, 16
    amplitude modulation (AM), 17–18
    combined modulation methods, 19–20
    frequency modulation (FM), 18–19
    phase modulation (PM), 19
  radio receivers, 33
    fundamental requirements, 34–6
    types, 36–9
  radio signals, 15
  radio wave propagation techniques:
    electromagnetic waves, properties of, 20
    free-space radiation, 20–3
  stages in, 15–16
Radio flexibility and capability, 115–18
RaDiO framework, 222–4
Radio frequency (RF) radiation, 137
  conveying intelligence, by radio waves, 144–6
  decibel, use of, 148–9
  frequency and wavelength, 143–4
  ionising and non-ionising radiations, 146
  radio transmission, history of, 137–8
  radio waves, nature of, 138–43
  terms used, explanation of, 146–8
Radio Link Protocol (RLC), 306, 308
Radio Network Controller (RNC), 306
Radio Resource Control (RRC) protocol, 308
Radio transmission, history of, 139–40
Radio wave propagation techniques:
  electromagnetic waves, properties of, 20
  free-space radiation, 20–3
Rake receiver, 310, 409
Random backoff, 385
Raptor codes, 209
'Rat-race' coupler, 66–7
Rate-compatible convolutional codes, 304
Rate compatible punctured convolutional (RCPC) codes, 205
Rate-distortion function, 201
Rate–distortion (R–D) optimization techniques, 229, 242
Rate–distortion optimized (RaDiO) streaming, 224
  basic framework, 224–6
  receiver-driven streaming, 226, 227
Ray methods, 319
Ray models, 324–7
  advanced techniques for, 327–32
  applications, 333–7
Reactances using transmission lines, 63–4
Real-time communication, 233
  architecture and fundamentals:
    media transport and control protocols, 236–9
    signaling protocols, 233–6
    systems, 233, 234
  quality of service:
    audio–video synchronization, 242–3
    available bandwidth, adapting to, 242
    combating losses, 241–42
    minimizing latency, 239–41
Real-time control protocol (RTCP), 238
Real-time transport protocol (RTP), 193, 236–8
Receive signal processors (RSPs), 76, 84
Receiver-driven streaming, 226, 227
Receiver selectivity, 34–5
Receivers, types of:
  double superheterodyne receivers, 39
  superheterodyne receiver, 37–9
  tuned radiofrequency receiver, 36–7
Red, green, and blue (RGB) color space, 470
  scRGB color space, 470
  sRGB color space, 469
Reduced function devices (RFD), 432
Reduced instruction set computers (RISCs), 96
Reed-Solomon codes, 208
Reflection coefficient, 43
Reflection transmission coefficients, 54
  mismatched loss, 57
  return loss, 56–7
  transmission coefficient, 57
  voltage reflection coefficient (Gv), in transmission lines, 55–6
  and voltage standing wave ratio, 57–8
  VSWR and reflection coefficient, 58
ReFlex, 7, 8
REGISTER method, 232
Release, 99, 306
Repeat-accumulate (RA) codes, 209
Request message, 234
Request to send (RTS) packet, 389, 417
Research into Advanced Communications in Europe (RACE), 262
Resource base class, 102
Resource Reservation Protocol (RSVP), 193–4
ResourceFactory class, 104
Response code, 235
Response message, 234
RESV messages, 195
Return loss, 56–7
RF and wireless technology, survey of:
  current technologies, 3
    applications, 8
    fixed location systems, 5–6
    local area systems, 6–7
    wide area mobile systems, 7–8

RF and wireless technology, survey of: (*continued*)
foreseeable future:
advanced networking, 10
increased access, 10–11
increased throughput, 10
smart antennas, 9–10
software radio, 8–9
ubiquity, 11
ultrawideband signals, 9
short history, 3
RF machines and RF plant, 148
RGB video signals, 464
RGB-YCbCr equations, 471–2, 472–3
Rich acknowledgments, 226–8
Ricochet system, 8
RIMAX algorithm, 345
Ring coupler, 66
RLC Protocol Data Units (RLC PDUs), 306
Robbing bits, 165
Rod/monopole antenna, 25
Round-robin (RR) scheduler, 301
Round-robin scheduling, 297–8, 300
RR control packet, 236
RT models, 325
RTCP SR control packets, 242

S-Video, 463
SAGE algorithm, 342–5
SAR (Segmentation And Reassembling) function, 171
scRGB color space, 470
SDTV:
RGB-YCbCr equations, 471–2
sample rate selection, 484
Second generation (2G) cellular systems, 71
Secure Sockets Layer (SSL), 401
Semiconductor devices, 140
Sender-based loss recovery technique, 241
Sender-driven streaming, 226
Sensitivity of radio receiver, 35–6
Service Level Agreements (SLAs), 168, 190, 199
Service-specific convergence sub-layer (SSCS), of AAL2, 174
Session Description Protocol (SDP), 236
Session initiation protocol (SIP), 233–4
Shannon's channel coding theorem, 204
Shannon's noiseless coding theorem, 202, 203
Shannon's source and channel theorems, 202
channel coding, 203–4
extensions, 205
source coding, 202
lossless source coding, 202
lossy source coding, 203
source-channel coding, 204–5
Shannon's source-channel coding theorem, 205
Shared Wireless Access Protocol (SWAP), 414
Shoe-box-sized Sensoria WINS NG sensors, 248
Short-circuit impedance, 51–2
Short interframe space (SIFS), 388, 416
Short message service (SMS), 8
Short range wireless applications and technologies, 413
Bluetooth, 425
error correction and encryption, 428–9
inquiry and paging, 429–30
packet format, 428
conflict and compatibility, 433
Bluetooth and Wi-Fi coexistence, improving, 435–6
ultra-wideband (UWB) technology, 436–9
wireless local area networks (WLAN), 413
HIPERLAN/2, 424
HomeRF Working Group, 414
network architecture, 415–17
physical layer, 417–24, 425
Wi-Fi, 414
Zigbee, 430
application, 433
architecture, 430–1
communication characteristics, 431–2
device types and topologies, 432
frame structure, collision avoidance, and reliability, 432–3
SIF (standard input format), 488
Signal dynamic range, 80
Signal processing, in software technology, 126–7
Signal propagation, on transmission lines, 52–3
Signal-to-noise ratios (SNRs), 36, 82, 214
Signal transfer points (STPs), 164
Signaling control agents, 196
Signaling protocols, 233
address, 235
header fields, 235–6
message body, 236, 237
Signaling system 7 (SS7), 163, 164
Silence suppression software, 191–92
Simple reflex agent, 125–6
SINCGARS (Single-Channel Ground and Airborne Radio System), 117
ASIP (Advanced System Improvement Program), 119
Sine wave, 138, 141
Single-bend attenuation method, 453, 455
Single-carrier receiver, 74, 75
Single-Frequency Network (SFN) operation, 293
Single loop antennas, 25–6
Single-mode (SM) fiber, 445, 447, 449
Slot line, 45
Slot time, 416
Smart agents models, 125, 126
Smart antennas, 9–10
SMPTE 267M, 486
SMPTE 274M, 495, 496
SMPTE 293M, 488
SMPTE 296M, 492
SNA (Systems Network Architecture), 185
Society of Motion Picture and Television Engineers (SMPTE), 483
Soft handover, 310
Software Assembly Descriptor (SAD), 107
Software capable radios, 116, 117–18
Software communications architecture (SCA), 101
application control, 106
application programming interface, 107–8

base components, 102–4
framework control, 104–6
profiles, 107
system control, 106–7
Software Component Descriptor (SCD), 107
Software defined radio (SDR), 69, 115, 116, 118, 121
applications and need for:
interception, 72
interoperability, 71–2
manufacturing platform, 72–3
architectures, 73
receiver, 73–7
transmit, 77–8
CDMA2000 and UMTS SDR receiver, case study of, 86–9
history and evolution, 70–1
implementation issues:
analog front end, 78–81
data conversion, 81
digital processing, 83–6
general converter requirements, 82–3
multiband radio, 70
multi-carrier/multi-channel radio, 70
multimode radio, 70
multi-rate radio, 70
as platform for cognitive radio, 91
applications, 109–10
cognitive waveform development, 111–12
development and design, 101–8
hardware architecture, 92–9
software architecture, 99–101
waveform development, 110–11
variable bandwidth, 70
Software-defined networks, *see* Virtual private networks (VPNs)
Software Package Descriptor (SPD), 107
Software radio, 8–9, 116
Software technology, for cognitive radio development, 125
artificial intelligence (AI) techniques, 125–6
networking protocols, 127
policy engines, 125
signal processing, 126–7
software communications architecture, 127–8
Source-channel coding, 204–5
Source coding, 202
lossless source coding, 202–3
lossy source coding, 203
Spatial compression, 509
Spatial multiplexing, 275
SpeakEasy, 70
Specialized private systems, 7–8
Specific absorption rate (SAR), 147
Spectral mask, 390
Spectrum awareness/frequency occupancy, 123
Spectrum awareness:
and potential for sublease/borrow, 128
Spectrum Policy Task Force (SPTF), 128
SR control packet, 238
sRGB color space, 469
SSRC, 235–6
Standard-definition video, 465
Standard input format (SIF), 509
Standards Developing Organizations (SDO), 263, 264
Standing wave ratio (SWR), 44
Start frame delimiter (SFD), 420
Strip line, 44–5
SUO-SAS (Small Unit Operations–Situational Awareness System), 121
Super high frequency (SHF) bands, 6
SuperFrames, 165
Superheterodyne receiver, 37–9
Switched trunking, 172
Switched Virtual Circuits (SVCs), 155
Symbol maximum a posteriori (MAP) rule, 202
Symmetric key system, 401
Synchronous connection oriented link (SCO), 426, 428, 429
Synchronous Optical Network (SONET), 167
Synchronous Residual Time Stamping (SRTS), 171
System Architecture Evolution (SAE), 259, 272
System-on-chip (SoC), 99
Systematic convolutional (RSC) codes, 214, 215

'Talk-over', 190
Tanner graph, 206, 207
Teledesic, 7
Temporal domain compression, 509–10
Temporal Key Integrity Protocol, 403
Terrestrial packet systems, 8
TestableObject parent class, 104
Third Generation Partnership Project (3GPP), 259, 261, 262–3
Time Division Duplex (TDD), 264, 265
Time Division Multiplexing (TDM), 154, 155, 167, 296, 298
architecture, 162
Time propagation delay, 43
Timestamp, 235
Tornado codes, 207
ToS (Type of Service), 183
Traditional telephony services, 268–9
Transcoding, 189
Transmission coefficient, 43, 57
Transmission lines, 41
characteristic impedance:
by measurement, 50–2
from primary electrical parameters, 48–50
couplers, 64
branch-line coupler, 65–6
'rat-race' coupler, 66–7
ring coupler, 66
as electrical components, 62
impedance relations, 63–4
input impedance, of low loss transmission lines, 63
reactances using transmission lines, 63–4
as transformers, 64
electrical transmission lines, types of, 44
coaxial and strip lines, 44–5
coupled lines, 45
microstrip and slot lines, 45
twin lines, 45
waveguides and coplanar waveguides, 44

Transmission lines (*continued*)
of finite length, 54
general properties, 42–4
line characteristic impedances and physical parameters, 45
coaxial line characteristic impedance, 46
microstrip line characteristic impedance, 46–8
twin parallel wire characteristic impedance, 46
propagation constant, 59–62
reflection transmission coefficients, 54
mismatched loss, 57
return loss, 56–7
transmission coefficient, 57
voltage reflection coefficient, 55–6
voltage standing wave ratio (VSWR), 57–8
VSWR and reflection coefficient, 58
signal propagation on, 52–3
typical commercial cable impedances, 52
waveform distortion and frequency dispersion, 53–4
Transmission policy, 224
Transmission Time Interval (TTI), 308
Transmit power control (TPC), 425
Transverse electromagnetic mode wave (TEM), 145
Transverse electromagnetic waves (TEM waves), 141
Tree-structured VQ, 504
Trellis-coded modulation (TCM), 211, 407
Tristimulus values, 478
TSG RAN, 262, 270
Tuned radio frequency receiver, 36–7
Turbo codes, 205, 214–15, 304
Twin lines, 45, 46
Twin parallel wire characteristic impedance, 46
Two way splitter, 32
Type of Service (ToS), 194

UDP (User Datagram Protocol) ports, 190, 193, 194
Ultrawideband (UWB) signals, 9
Ultra-wideband (UWB) technology, 407, 436–9
UMTS SDR receiver, 86–9
Unbalanced antenna, 29
Unbalanced line, 30
Uniform Theory of Diffraction (UTD), 319
Universal Mobile Telecommunications Services (UMTS), 260
Unlicensed National Information Infrastructure (UNII) band, 396
Unmanned aerial vehicles (UAVs), 252
Unmatched lines, 54
Unspecified bit rate (UBR), 168
Uplink scheduling, 299–300
Upper logical link control layer, 384
User agent client (UAC), 234
User agent server (UAS), 234
User Datagram Protocol (UDP), 193
Utility-based agent, 126
UTRA (Universal Terrestrial Radio Access), 264
3GPP specifications for, 263
UWB (Ultra Wide Band) systems, 317
propagation modelling by ray tracing, 332
UWB-SAGE algorithm, 343

Vanu Anywave™ BaseStation, 118
Variable bandwidth, 70
Variable bit rate (VBR), 168, 174
Variable length coding (VLC), 503, 509
Varying queuing delay, 238
Vertical sync, 464
Very small aperture terminal (VSAT) systems, 6, 8
Video, introduction to:
analog vs. digital, 463
application block diagrams, 466
Digital media adapters, 467
Digital television set-top boxes, 467
DVD player, 466–7
audio and video compression, 465–6
resolution, 465
timing, 464–5
video data, 463–4
Video compression, 501
application, 505–6
closing remarks, 521
digital video signals and formats, 506
analog video signals, sampling of, 506–8
digital video formats, 508–9, 510
techniques, 509
discrete cosine transform (DCT), 511–13
entropy and predictive coding, 510–11
motion compensation and estimation, 514–16
quantization, 513–14
video encoding standards and H.261, 516
H.261 video encoder, 517–21
Video encoding standards, 516
and H.261 video encoder, 517–21
Video payload format, 238–9
Video redundancy coding (VRC), 240
Video signals, 483
1080i and 1080p systems:
interlaced analog component video, 493, 495
interlaced digital component video, 495–6
progressive analog component video, 495
progressive digital component video, 496–7
480i and 480p systems:
interlaced analog component video, 484–5
interlaced analog composite video, 484
interlaced digital component video, 486–8
progressive analog component video, 485–6
progressive digital component video, 488
SIF and QSIF, 488
576i and 576p systems:
interlaced analog component video, 488–9
interlaced analog composite video, 488
interlaced digital component video, 490
progressive analog component video, 489–90
progressive digital component video, 490–2
720p systems:
progressive analog component video, 492
progressive digital component video, 492–3, 494
digital component video background, 483
coding ranges, 483
EDTV sample rate selection, 484

HDTV sample rate selection, 484
SDTV sample rate selection, 484
video systems, 498
Virtual carrier sense mechanism, 417
Virtual circuit (VC), 169–70
Virtual machine (VM), 109
Virtual private network (VPN), 8, 197–8, 200, 401
Viterbi algorithm, 212
Viterbi trellis decoder, 395
VLANs, 185, 194
Voice Activity Detection (VAD), 191–92
Voice Frame Relay access devices (VFRADs), 175, 176, 177
Voice networking, 165–6
Voice over ATM (VoATM), 167
addressing, 169
ATM trunking, using AAL2, 174–5
benefit, 170
circuit emulation services (CES), 173–5
and delay, 170
routing, 169–70
signaling, 169
standards and specifications, 172–3
technical challenges, 170
compression algorithms, 171
echo, 170
end-to-end delay, 170–71
signaling, 171
silence suppression, 171
synchronization, 171–72
Voice over Frame Relay (VoFR), 168, 175
network equipping, 175
controlling variable delay, 177–8
data network and implications, 176
data network, voice-enabling, 176–7
echo cancellation, 178
fragmentation, 177
prioritization, 177
silence suppression, 178
technology, 177
voice compression, 178
voice switching requirements, 178–9
Voice over IP (VoIP), 168, 179
bandwidth, 179, 182–3
delay, 181
jitter, 184
latency, 185–6
manageability, 183–4
market size, 180–82
new services, 179–81
packet loss, 181–82
progressive deployment, 180
quality of service (QoS), 186–7
technology, *see* VoIP technology
voice packet prioritization, 184–5
Voice services, 8
Voice switching, 172–3, 175–6
Voice traffic characteristics, 160
convergence technologies, 162–4
data, addition of, 161–62
TDM architecture, 162
Voice trunking, 172
VoIP technology, 189
frame relay, 198
IP network protocols, to support voice, 193
class-based queuing (CBQ), 196–7
Class of Service (CoS) vs. Quality of Service (QoS), 194
differential services, 194
IEEE 802.1 p/Q, 194
prioritized queuing, 196
priority queuing, 196
Resource Reservation Protocol (RSVP), 195–6
UDP ports, 194
VLANs, 194
weighted fair queuing (WFQ), 196
IP network, voice traffic needs in, 189
codec selection, 192
duplex, 192
echo, 191
jitter, 190
packet delay, 190
packet loss, 190–92
packet mis-order, 191
silence suppression software, 191–92
transcoding, 191
IP softphone, PC considerations using, 192–3
Network Address Translation (NAT), 198
network design recommendations, 198–200
VPN (Virtual Private Network), 197–8
WAN considerations, 197
Voltage reflection coefficient, in transmission lines, 55–6
Voltage standing wave radio (VSWR), 57–8, 92
and reflection coefficient, 58
Voronoi region, 210

Wave impedance, 22
Waveform development, in SDR, 110–11
Waveform distortion and frequency dispersion, 53–4
Waveguides, 44
Web Ontology Language (OWL), 125
Weighted Fair Queuing (WFQ), 196, 197
Wi-Fi devices, 7, 413, 414, 434, 435–6
Wi-Fi PHY (802.11b), 394–6
Wi-Fi Protected Access (WPA), 403
Wide area mobile systems, 4, 7
cellular, 8
mobile satellite services, 7
paging systems, 7
specialized private systems, 7–8
terrestrial packet systems, 8
Wide area networks (WANs), 157, 160, 190, 384
Wideband code division multiple access (WCDMA), 260–1, 271, 280
evolution, 305–11
overall architecture, 306–8
physical layer, 308–10
resource handling and packet-data session, 311
Wideband fading prediction, 369–70
Wideband short-term fading, 356–9

Wider bandwidth, 280–82
Wired equivalent privacy (WEP), 401, 402, 414
Wireless communication, short history of, 3
Wireless local area networks (WLANs), 383, 413
  802.11 WLANs, 386
    802.11 (In) Security, 401–3
    802.11 alphabet soup, 392–4
    802.11 architecture, 387, 388
    802.11 classic direct-sequence PHY, 389–92, 393
    802.11a PHY, 396–9, 400
    802.11g PHY, 399–400
    MAC and CSMA/CA, 387–9
    Wi-Fi PHY (802.11b), 394–6
  HiperLAN and HiperLAN, 2, 403–4
  HIPERLAN/2, 424
  HomeRF Working Group, 414
  from LANs, 385–6
  LANs to PANs operation, 404
    802.15.3 task group, 406–7
    Bluetooth, 404–6
    UWB PANs, 407–11
  network architecture, 415–17
  physical layer, 417
    802.11a and OFDM, 423–4
    802.11b, 420–3
    DSSS PHY, 419–20, 421
    FHSS PHY, 418–19
    IEEE 802.11 basic, 418
    Infrared PHY, 418
  Wi-Fi, 414
Wireless local area networks (WLANs), 7
Wireless PANs (WPANs), 383
Wireless sensor networks, 247
  advantages, 250
    detection advantage, 251
    energy advantage, 250–51
  application, 251
    habitat monitoring, 251–53
    smart transportation, 254–5
    tracking chemical plumes, 253
  collaborative processing, 255
  constraints and challenges, 247–50
  key definitions, 255–6
wxWidgets, 110

xvYCC color space, 474

Yagi-Uda array, 28–9
YCbCr color space:
  4:4:4 YCbCr format, 473–4, 476
  RGB-YCbCr equations, 471–3
YCbCr video signals, 464
YIQ color space, 471
YPbPr color space, 463
YUV color space, 470–1

Zigbee, 430
  application, 433
  architecture, 430–1
  communication characteristics, 431–2
  device types and topologies, 432
  frame structure, collision avoidance, and reliability, 432–3

Information on source books

Erik Dahlman, Stefan Parkvall, Johan Sköld & Per Beming
3G Evolution: HSPA and LTE for Mobile Broadband, 2nd Edition
9780123745385
Aug 2008
£69.99/$119.95
The only book with the latest 3G technologies to achieve the next key stage towards mobile broadband
* Written by engineers working closely in 3GPP
* Includes details of the standards and technologies (160 new pages): LTE radio interface architecture, LTE physical layer and LTE access procedures
* Gives clear explanations of the role of OFDM and MIMO technologies in HSPA and LTE
Academic Press

Bruce Fette
Cognitive Radio Technology
9780750679527
2006
£51.99/$87.95
The first comprehensive book on cognitive radio technology - by one of its pioneers
* The first book on a revolutionary technology that will be critical to military, emergency, and public safety communications
* A multi-contributed volume written by the leaders in this exciting new area
* Describes the location-determination capabilities of cognitive radio (the precise location of all units in a cognitive radio network can be determined in real time)
Newnes

Casimer DeCusatis
Handbook of Fiber Optic Data Communication
9780122078910
2002
£82.00/$180.00
The leading reference for optical fiber technology, networking, protocols, applications, and manufacturing.
* Includes emerging technologies such as Infiniband, 10 Gigabit Ethernet, and MPLS Optical Switching
* With an expanded listing of references on the World Wide Web, plus hard-to-find references for international, homologation, and type approval requirements
* Convenient tables of key optical datacom parameters and glossary with hundreds of definitions and acronyms
Academic Press

Alan C. Bovik
Handbook of Image and Video Processing
9780121197926
2005
£76.00/$142.00
THE must-have reference for both students and practitioners of image and video processing
* No other resource for image and video processing contains the same breadth of up-to-date coverage
* Each chapter written by one or several of the top experts working in that area
* Includes all essential mathematics, techniques, and algorithms for every type of image and video processing used by electrical engineers, computer scientists, internet developers, bioengineers, and scientists in various, image-intensive disciplines
Academic Press

Edited by Farid Dowla
Handbook of RF and Wireless Technologies
9780750676953
2004
£68.99/$119.00
Industry experts have contributed to this unique handbook that covers the world of cutting-edge RF and wireless design
* A comprehensive survey of current RF and wireless engineering practice
* Heavy emphasis on practical applications and design guidelines
* Multiple contributors assure a wide range of perspectives and avoids individual bias
Newnes

Luis Correia
Mobile Broadband Multimedia Networks: Techniques, Models and Tools for 4G
9780123694225
2006
£51.99/$97.95
10. A complete reference of the latest technologies for achieving 4th generation mobile and wireless communications
* Based on the well known and respected research of the COST 273 project 'Towards Mobile Broadband Multimedia Networks', whose previous models have been adopted by standardisation bodies such as ITU, ETSI and 3GPP
* Gives methods, techniques, models and tools for developing 4th generation mobile and wireless communication systems
* Includes the latest development of key technologies and methods such as MIMO systems, ultra wide-band and OFDM
Academic Press

Mihaela van der Schaar & Philip A. Chou
Multimedia over IP and Wireless Networks
9780120884803
03/2007
£49.99/$79.95
A focused presentation of online multimedia content delivery problems and their solutions!
* Up-to-date coverage of existing standards for multimedia networking
* Synergistic tutorial approach reinforces knowledge gained in previous chapters
* Balanced treatment of audio and video with coverage of end-to-end systems
Academic Press

Ronald Kitchen
RF and Microwave Radiation Safety
9780750643559
2001
£74.00/$119.00
A practical handbook for all involved in electronic design and safety assessment, including the use of measuring instruments and methods, radiation hazards and risks resulting from electromagnetic interference
* Leading professional guide to RF and microwave safety issues
* Covers radiation hazards and risks resulting from electromagnetic interference
* Revised in line with the latest the EU and international standards
Newnes

Daniel M. Dobkin
RF Engineering for Wireless Networks
9780750678735
2005
£48.99/$88.95
The Essential One-Volume Reference for the Design and Engineering of all Types of Wireless Networks
* The A-to-Z guide to wireless network engineering—covers everything from basic electromagnetic theory to modulation techniques to network planning and implementation
* Engineering and design principles covered are applicable to any type of wireless network, including 802.11, 802.16, 802.20, and Bluetooth
* Discusses state-of-the-art modulation techniques such as ultra wideband (UWB) and orthogonal frequency-division multiplexing (OFDM)
Newnes

Alan Bensky
Short-range Wireless Communication
9780750677820
2004
£43.99/$68.95
This is a complete "tool kit" for the hottest area in wireless design engineering today
* No other book contains EVERYTHING pertaining to short-range wireless design
* Covers all the hot topics like 802.11, Zigbee, Wi-Fi and Bluetooth
* Valuable CD-Rom included that will simplify the design process
Newnes

Keith Jack
Video Demystified
9780750683951
06/2007
£39.99/$69.95
The market leader, this book is now updated with info on today's hottest technology
* The next generation of digital video - Blu-ray and HD-DVD thoroughly introduced
* All broadcast and satellite standards completely updated
* Essential information for the upcoming transition of television signals from analog to digital
Newnes

Juanita Ellis, Charles Pursell & Joy Rahman
The Convergence of Voice, Video & Network Data
9780122365423
2003
£45.99/$76.95
The first book that covers convergence of voice, video and data networks from a vendor-neutral perspective
* Covers the latest standards and those being developed in an ever-evolving field
* Provides insight into the latest technology of video and data over wireless networks and how convergence will be a driving force in this industry
* Provides an understanding of the true capabilities behind each vendor's solution to allow for informed buying decisions
Academic Press

Feng Zhao & Leonidas Guibas
Wireless Sensor Networks: An Information Processing Approach
9781558609143
2004
£48.99/$73.95
Introduces fundamental issues and technology constraints of sensor networks in a concrete way
* Written for practitioners, researchers, and students and relevant to all application areas, including environmental monitoring, industrial sensing and diagnostics.
* Skilfully integrates the many disciplines at work in wireless sensor network design, including signal processing and estimation and communication theory and protocols
* Demonstrates how querying, data routing, and network self-organization can support high-level information-processing tasks
Morgan Kaufman

Ron Olexa
Implementing 802.11, 802.16, and 802.20 Wireless Networks
9780750678087
2001
£33.99/$68.95
Be on the cutting-edge of the wireless revolution! Master planning, implementing and troubleshooting wireless networks with the newest 802 standards
* Covers the widely used 802.11 family, as well as the new 802.16 and 802.20 standards
* Focuses on big commercial network implementations, such as in public buildings and businesses
* Author has over 25 years of experience with cellular systems and wireless networks
Newnes
33% Discount Code: 93641
http://www.elsevierdirect.com/product.jsp?isbn=9780750678087

Claude Oestges & Bruno Clerckx
MIMO Wireless Communications
9780123725356
05/2007
£59.99/$89.95
The first book on MIMO systems to bring together propagation and signal processing - essential for exploiting the real potential of this very promising technology
* Presents space-time coding techniques for real-world MIMO channels
* Contains new design methodologies and criteria that guarantee the robustness of space-time coding in real life wireless communications applications
* Evaluates the performance of space-time coding in real world conditions
Academic Press
33% Discount Code: 93641
http://www.elsevierdirect.com/product.jsp?isbn=9780123725356

Wing-Kuen Ling
Nonlinear Digital Filters
9780123725363
05/2007
£39.99/$79.95
The only book to give practical design techniques for implementing high quality nonlinear digital filters - a key task in signal processing
* Practical techniques to design nonlinear digital filters- essential in signal processing for communications engineering
* Includes practical, 'real-world' applications of nonlinear digital filter design
* Many examples, together with MATLAB source code, are available on a companion website
Academic Press
33% Discount Code: 93641
http://www.elsevierdirect.com/product.jsp?isbn=9780123725363

David Morgan
Surface Acoustic Wave Filters
9780123725370
07/2007
£59.99/$99.95
This book gives the fundamental principles and device design techniques for surface acoustic wave filters
* A thorough revision of a classic on surface acoustic wave filters first published in 1985 and still in print
* Uniquely combines easy -to -understand principles with practical design techniques for all the devices in widespread use today
* Includes a new foreword by Sir Eric Albert Ash
Academic Press
33% Discount Code: 93641
http://www.elsevierdirect.com/product.jsp?isbn=9780123725370